

ChemExplain: A Multimodal Benchmark for Explanatory Chemical Reasoning

Aotian Chen¹, Frazier N. Baker¹, Reza Averly¹, Shan Zhang²,
Tianao Chen¹, Daniel Adu-Ampratwum¹, Xia Ning¹, Yuan Xue¹,✉

¹ The Ohio State University

² Adelaide University

✉ Correspondence: xue.643@osu.edu

Abstract

Current chemistry benchmarks primarily evaluate outcome prediction or reaction pathway generation, but rarely test whether a model can explain why an observed chemical effect occurs. We introduce ChemExplain, a benchmark of 1,180 multiple-choice chemistry questions drawn from 383 peer-reviewed papers. Each question pairs a molecular or reaction image with an observed chemical effect and four candidate explanations, each associated with a different structural feature visible in the image. Answering requires the model to perceive the depicted structure, apply chemical knowledge, and identify which feature most directly explains the effect, a capability we term explanatory attribution. The correct feature-mechanism relation is derived from supporting evidence in the source paper rather than invented by the question-generation LLM. We construct the benchmark automatically with a multi-agent framework that applies progressive quality control at three stages and validate it with eight chemistry domain experts. The strongest evaluated model reaches 78.6% accuracy, leaving considerable room for improvement. On a 600-question expert-reviewed subset, reasoning-trace analysis classifies 71.4% of the strongest model’s errors as attribution failures. Attribution is also the largest error category for the open-source and chemistry-specific models included in our analysis.

They rarely ask whether a model can identify which visible structural feature most directly explains an observed chemical effect. A model may predict the correct product while relying on the wrong chemical rationale. This explanatory capability matters because observed effects guide which structural choices to preserve, modify, or discard. Such decisions arise in molecular design, mechanism analysis, and experiment planning, where attributing an effect to the wrong feature leads to modifying the wrong part of a molecule.

Answering such questions requires a multi-step reasoning chain. The model must perceive the relevant structural features in a molecular or reaction image, apply the chemical knowledge needed to relate those features to the observed effect, and determine which feature is primarily responsible in the specific structural context. We refer to this final decision as *explanatory attribution*. Existing benchmarks probe individual parts of this process, including structure recognition (Guo et al., 2024), chemical question answering (Guo et al., 2023; Huang et al., 2024), and outcome prediction (Li et al., 2025a; Xu et al., 2025), but they do not evaluate the complete process in a single task (Appendix A). They therefore cannot fully distinguish whether a model fails because it misreads the structure, lacks the relevant chemistry, or attributes the effect to the wrong feature.

Studying this capability requires a benchmark grounded in scientific literature, but the necessary evidence is difficult to recover. Chemical structures appear in figures, while the explanation connecting a structure to an observed effect may span captions and body text. Existing extraction tools (Swain and Cole, 2016; Fan et al., 2024; Leong et al., 2025; Shah et al., 2025) retrieve structures and associated text but do not determine whether the retrieved text contains a structure-to-effect explanation. We therefore develop a multi-agent framework that isolates individual structures, retrieves and validates ex-

1 Introduction

Vision-language models (VLMs) are increasingly used in chemistry for molecular design, retrosynthesis, and reaction optimization, where success depends on reasoning over molecular structures, reaction schemes, and chemical properties. Existing chemistry benchmarks evaluate many of these capabilities (Mirza et al., 2025; Guo et al., 2023; Huang et al., 2024; Zhao et al., 2025), but primarily test reaction outcome prediction or pathway generation (Li et al., 2025a; Xu et al., 2025) (Fig. 1).

planatory evidence, and applies quality control after each stage. Using this framework, we construct ChemExplain from about 5,000 chemistry papers while retaining the source paper, figure, structure, and supporting evidence for each selected question.

ChemExplain contains 1,180 multiple-choice questions drawn from 383 selected peer-reviewed papers. It covers three reasoning types: how a structural feature produces an observed effect (Mechanistic), why a feature was introduced to achieve a target property (Design), and why a structural feature requires particular reaction conditions (Condition) (Appendix Fig. 5). Each question presents a molecular or reaction image, an observed effect, and four explanations associated with different visible structural features. The model must select the explanation that best accounts for the effect in the depicted structure (Appendix Fig. 6). The correct feature-mechanism relation is derived from supporting evidence in the source paper rather than invented by an LLM. Benchmark quality is assessed through LLM-as-a-Judge screening and independent evaluation by eight chemistry domain experts.

We evaluate 12 proprietary, open-source, and chemistry-specific VLMs on ChemExplain. The strongest model reaches 78.6% accuracy. To locate where models fail, we manually classify the reasoning traces of three representative models on a stratified expert-reviewed subset as perception, knowledge, or attribution failures. Attribution accounts for 71.4% of the strongest model’s errors (Fig. 4) and is the largest category for the open-source and chemistry-specific models, although they make more perception and knowledge errors.

Our contributions are threefold. (1) We introduce ChemExplain, a literature-grounded benchmark of 1,180 multiple-choice questions for explanatory attribution in chemistry, independently evaluated by eight chemistry domain experts. (2) We present a multi-agent construction framework with progressive quality control and item-level provenance linking each accepted question to its source-paper evidence. (3) We systematically analyze model failures across perception, chemical knowledge, and explanatory attribution, finding that attribution remains the largest error category for all models.

2 Related Work

Chemistry Benchmarks. Existing chemistry benchmarks mostly test knowledge, prediction, or

Benchmark	Size	Modality	Level	Curation	Source	Attr.
GPQA (Rein et al., 2024)	448	Text	Grad	Manual	Expert	×
ChemLLMBench (Guo et al., 2023)	800	Text	Mixed	Semi	Expert	×
ChemBench (Mirza et al., 2025)	2,788	Text	UG-Grad	Semi	Expert	×
ChemCoTBench (Li et al., 2025a)	1,495	Text	Mixed	Semi	Expert	×
oMeBench (Xu et al., 2025)	2,871	Text	Mixed	Semi	Expert	×
SUPERChem (Zhao et al., 2025)	500	Multimodal	Mixed	Manual	Expert	×
MolTextQA (Laghvarapu et al., 2025)	500K	Multimodal	Mixed	Auto	LLM	×
ChemLitQA (Wellawatte et al., 2025)	1,411	Text	Grad	Semi	LLM	×
ChemExplain	1,180	Multimodal	PhD	Auto	Literature	✓

Table 1: ChemExplain compared with existing chemistry benchmarks. UG = undergraduate, Grad = graduate. **Source**: origin of the correct answer. **Attr.**: whether the task requires explanatory attribution.

retrieval, and many are text-only. General scientific benchmarks such as GPQA (Rein et al., 2024) and SciBench (Wang et al., 2023) include chemistry questions but do not involve molecular diagrams or reaction schemes. Chemistry-specific benchmarks cover property prediction, structure elucidation, and name conversion (Guo et al., 2023; Mirza et al., 2025; Huang et al., 2024; Guo et al., 2024), while ChemCoTBench (Li et al., 2025a), oMeBench (Xu et al., 2025), and SUPERChem (Zhao et al., 2025) test reaction outcome prediction and pathway generation. Literature-based QA benchmarks (Laghvarapu et al., 2025; Amiri and Bocklitz, 2025; Chen et al., 2024; Wellawatte et al., 2025) derive questions from papers but mainly test factual retrieval. These benchmarks rarely ask which visible structural feature explains an observed effect. ChemExplain targets this capability (Tab. 1).

Benchmark Construction. Benchmarks also differ in where the correct answer comes from. Some literature-based QA benchmarks use LLMs to generate reference answers, introducing a risk of unsupported explanations when those answers are not independently verified (Wellawatte et al., 2025). Agentic pipelines such as AutoConverter (Zhang et al., 2025) and MicroVQA (Burgess et al., 2025) instead convert human-written question-answer pairs into multiple-choice questions. Deriving the answer from a paper requires locating the text that explains the effect, which extraction tools (Swain and Cole, 2016; Fan et al., 2024; Leong et al., 2025; Shah et al., 2025) do not separate from procedural or descriptive text. ChemExplain locates and validates this explanatory evidence and derives each correct answer from it (Appendix A).

3 Method

Our goal is to construct benchmark questions from chemistry papers that ask which visible structural feature most directly explains an observed chemical effect. Each question pairs a molecular or reaction

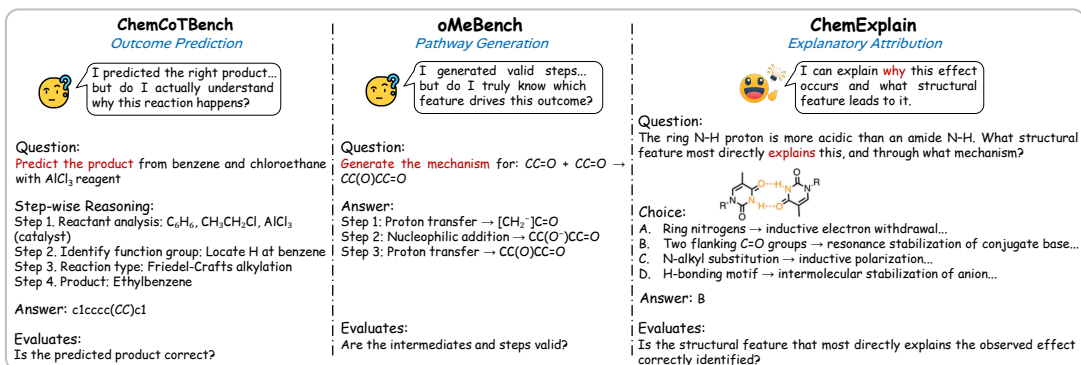


Figure 1: Task-level comparison with representative chemistry benchmarks. ChemCoTBench and oMeBench evaluate reaction outcome prediction and pathway generation, respectively. ChemExplain instead evaluates whether a model can identify which visible structural feature most directly explains an observed chemical effect.

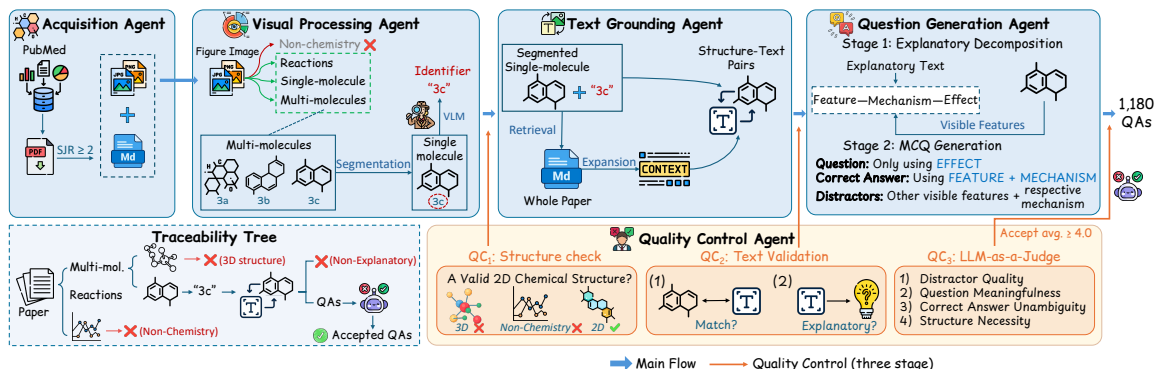


Figure 2: Overview of the multi-agent benchmark construction framework. Four specialist agents perform acquisition, visual processing, text grounding, and question generation, with blue arrows indicating the main data flow. Bottom left: a traceability tree shows how one source paper branches into candidate structures and how invalid candidates are filtered out across stages. Bottom right: a dedicated quality control agent applies three progressive checks between stages: QC₁ validates 2D chemical structures, QC₂ verifies entity match and explanatory evidence, and QC₃ screens generated questions on four quality dimensions.

image with an effect and a set of candidate explanations, and remains traceable to the supporting evidence in the source literature.

Figure 2 summarizes the multi-agent construction framework. Four specialist agents perform paper acquisition, visual processing, text grounding, and question generation. The acquisition agent converts PDFs into structured text and figure images with MinerU (Wang et al., 2024). The visual processing agent identifies chemistry figures, segments them into molecular or reaction instances, and recovers paper-specific identifiers such as "compound 4b." The text grounding agent uses each identifier to retrieve evidence and determine whether it contains a structure-to-effect explanation. The question generation agent converts validated structure-evidence pairs into multiple-choice items while preventing answer leakage.

A quality control agent checks the output after each core stage. QC₁ removes invalid or non-2D chemical structures. QC₂ verifies entity match

and explanatory value. QC₃ screens each generated item on four quality dimensions (Section 4.3). The provenance tree records each candidate and its quality-control decisions, linking every accepted question to its source paper, figure, structure, and supporting evidence. Sections 3.1–3.3 describe the three stages in detail, with additional implementation information in Appendix C.

3.1 Visual Processing

Chemistry figures often contain multiple molecular diagrams or reaction schemes, whereas each benchmark item must be linked to one molecule or reaction. The visual processing stage therefore isolates individual structures and recovers the identifiers that connect them to the article text.

A ResNet-50 classifier (He et al., 2016) separates chemistry figures from spectra, plots, and experimental images. DECIMER Segmentation (Rajan et al., 2021) isolates structures in multi-molecule figures, while RxnScribe (Qian et al., 2023a) de-

composes reaction schemes. We retain the segmented structures as images because SMILES conversion disproportionately removes complex structures, including many cases with detailed mechanistic discussion (Section 5.4). Existing co-reference tools such as MolCoref (Fan et al., 2024) recover identifiers for only a minority of our figures. We instead highlight each target structure in the original figure and use GPT-4o-mini to recover its identifier. Section 5.4 compares these alternatives.

3.2 Text Grounding

From the source literature corpus, we observe that most textual mentions of a molecule are procedural or descriptive rather than explanatory. Thus, we require evidence that connects a visible structural feature to an observed effect. GPT-4o-mini searches the full paper using the structure identifier, image, caption, and in-text figure references. Because the effect and mechanism often appear in adjacent sentences, each retrieved sentence is expanded with semantically related sentences from the same paragraph. Section 5.4 quantifies the contribution of this context expansion.

QC₂ evaluates both *entity match* and *explanatory value*. Entity match indicates whether the passage refers to the target molecule or reaction. Explanatory passages state how a structural feature produces the observed effect. Partially explanatory passages identify the feature but omit the mechanism, while non-explanatory passages are procedural or descriptive. We discard entity mismatches and non-explanatory passages. Partial explanations are retained only as context, and a structure advances to question generation only if its evidence block contains an explanatory sentence.

3.3 Question Generation

The question generation agent converts each validated structure-evidence pair into a multiple-choice question. The source paper determines the correct feature-mechanism relation, while GPT-5.1 reformulates this evidence into the question and answer options. The model is not allowed to introduce a correct-answer mechanism that is absent from the supporting evidence. The distractors use other genuine features visible in the same structure.

Two-stage generation. GPT-5.1 first decomposes the explanatory evidence into three fields: (1) Effect, the observed outcome without naming its structural cause (e.g., “resists oxidative degradation”); (2) Feature, the feature identified by

the source paper as primarily responsible; and (3) Mechanism, how that feature produces the effect. The question stem is generated from Effect alone, so Feature and Mechanism cannot leak into the wording of the stem. The correct option restates the Feature and Mechanism fields. Each distractor pairs another visible structural feature with a chemically plausible mechanism for the same effect.

Anti-shortcut constraints. We apply two constraints to reduce textual shortcuts without reading the image content. *Same-effect competition* requires every option to explain the same observed effect, reducing the possibility of identifying the correct option through simple textual matching. *Matched confidence* requires comparable tone and assertiveness across options and excludes hedging terms that could reveal distractors. Answer positions are randomized. Appendix H provides the complete prompts.

4 ChemExplain Construction and Validation

4.1 Source Corpus and Construction Yield

To instantiate the framework at scale, we draw source papers from the PubMed Central Open Access Subset (PMC OA), which provides articles under licenses that permit text mining and derivative use. We query PMC for papers in organic synthesis, materials chemistry, catalysis, and medicinal chemistry, yielding approximately 24,000 candidate papers. We further restrict the source corpus to journals with SCImago Journal Rank (SJR) ≥ 2.0 (SCImago, 2024), leaving approximately 5,000 papers in the final source corpus. Fig. 3 summarizes the construction yield at each stage.

From this filtered corpus, the acquisition agent extracts 73,536 figures. After figure classification, segmentation, identifier extraction, and structure validation, the visual processing agent yields 24,302 valid molecular or reaction instances spanning 2,906 papers. The text grounding agent retrieves and validates explanatory evidence for these instances: 9,670 (39.8%) yield retrieved text, and 1,264 of these (13.1%) pass both entity-match and explanatory-value validation, which qualifies the instance for question generation. The question generation agent converts these 1,264 candidates into questions, of which 1,180 (93.4%) pass the quality control agent’s LLM-as-a-Judge screening (Section 4.3) and are accepted into the final benchmark. Section 5.4 analyzes where these losses occur.

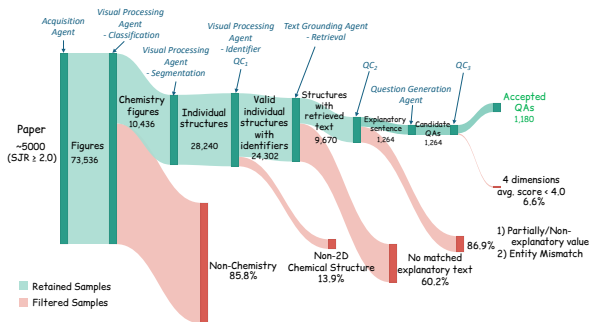


Figure 3: Construction yield and attrition from $\sim 5,000$ source papers to 1,180 accepted questions. Green denotes retained samples, red denotes filtered samples annotated with rejection cause and loss percentage, and stage labels indicate the responsible agent.

Dimension	$\bar{H}$	AC ₂			MAE $\downarrow$
		H-H	H-L	SE	
Distractor Quality	4.50	0.909	0.948	0.009	0.497
Meaningfulness	4.42	0.898	0.936	0.009	0.505
Answer Unambiguity	4.43	0.899	0.929	0.011	0.407
Struct. Necessity	4.75	0.946	0.861	0.022	0.785
Avg. (Score)	4.53	0.913	0.919	0.013	0.549

Table 2: LLM-as-a-Judge reliability ($n=600$ stratified sample from accepted questions, 8 human experts). $\bar{H}$: human expert mean; AC₂: Gwet’s AC₂; H-H: human inter-rater; H-L: human-LLM; SE: analytic standard error of the H-L estimate; MAE: mean absolute error.

4.2 Dataset Statistics

The final benchmark contains 1,180 accepted questions drawn from 383 peer-reviewed papers (Appendix Fig. 5). The three reasoning categories are: *Mechanistic*, which asks how a structural feature leads to an observed effect (55.9%); *Design*, which asks why a feature was introduced to achieve a target property (39.7%); and *Condition*, which asks why a structural feature requires specific reaction conditions (4.3%). In terms of entry type, the benchmark contains 338 molecule questions (28.6%) and 842 reaction questions (71.4%). These proportions follow the source literature, in which explicit structure-to-condition explanations are uncommon. Each paper contributes 3.08 questions on average (median 3, maximum 8). Appendix J shows examples of each category and entry type.

The benchmark is comparable in size to expert-curated reasoning benchmarks such as GPQA (448) and SUPERChem (500), and its automated construction supports extension to new source papers and subdomains without manual question writing.

4.3 Quality Validation

We evaluate benchmark quality through a two-phase protocol that combines full-corpus LLM screening with expert assessment of a stratified

subset. Human experts and the LLM judge independently rate the same 600 questions using the same four quality dimensions. Their agreement measures how closely the automated screening reproduces expert quality judgments.

LLM-as-a-Judge Screening. We use GPT-5.2 as the judge model with explanation-first prompting (Zheng et al., 2023; Tan et al., 2024), so that the model first analyzes each dimension before assigning a numeric score. The judge evaluates all 1,264 candidate questions on a 1–5 Likert scale along four dimensions: *Distractor Quality*, whether incorrect options remain plausible without careful analysis of the depicted structure; *Question Meaningfulness*, whether the question probes an explanatory relationship rather than superficial recall; *Correct Answer Unambiguity*, whether the labeled answer is the best-supported explanation and no competing option is equally defensible; and *Structure Necessity*, whether solving the question requires the depicted structure rather than the question text alone. To reduce answer-order effects (Shi et al., 2025), each question is scored three times with shuffled option order, and the scores are averaged. Of the 1,264 candidates, 1,180 (93.4%) achieve an average score of at least 4.0 and are accepted. The remaining 84 are rejected (Appendix E.5 analyzes the rejections and Appendix E.7 shows examples).

Human Expert Evaluation. Eight chemistry experts, including five faculty members and three PhD or postdoctoral researchers, independently evaluate a stratified sample of 600 accepted questions. The sample contains 300 molecule questions and 300 reaction questions, includes all 51 Condition questions together with 288 Mechanistic and 261 Design questions, and uses different source papers whenever possible. All eight experts rate every question on every dimension, yielding 4,800 expert-question ratings per dimension (Tab. 2). Each evaluator first answers the question without access to the source-paper text and then reviews the supporting evidence before assigning quality scores on the same 1–5 scale (Appendix I). In this blind phase, at least five of the eight experts select the labeled answer for 94.0% of questions (Appendix D.1).

Agreement and Validation Results. We compare judge and expert scores to determine whether the judge and the experts apply the quality rubric consistently. The mean judge scores on the expert subset are close to those on the full accepted benchmark after Holm–Bonferroni correction, with absolute differences of 0.003–0.125 points across

dimensions (Appendix E.2). We report Gwet’s AC₂ with quadratic weights (Gwet, 2014) because ratings cluster near the top of the scale, where weighted κ and Krippendorff’s α suffer from the prevalence paradox, together with MAE and confidence intervals (Klie et al., 2024; Appendix E.1).

Tab. 2 summarizes the results. The experts assign a mean quality score of 4.53 across dimensions, and every question also passes the 4.0 acceptance threshold under expert scores. Human-human agreement is high, with an average AC₂ of 0.913, indicating that the rubric is applied consistently across experts. Human-LLM agreement is similar, with an average AC₂ of 0.919 and all four dimensions above 0.86. The judge also agrees with the experts on rejected candidates: on 60 rejected questions that the experts rate without knowing the judge’s decision, both assign a mean score of 3.68 (Appendix Tab. 17). These results support using the judge as a scalable quality-control stage for all 1,264 candidates. Claude-Opus-4-7 and Gemini-3.1-Pro yield comparable or higher agreement on the same subset (Appendix E.3), showing that the result is not specific to the GPT judge.

5 Experiments and Analysis

5.1 Main Results

We evaluate ChemExplain on a diverse set of vision-language models, including closed-source systems (Gemini-3.1-Pro (DeepMind, 2026), GPT-5.4, GPT-5, GPT-4o (Singh et al., 2025; Achiam et al., 2023), Claude Sonnet 4.6 (Anthropic, 2025)), open-source models (Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), Phi-4-reasoning-vision (Aneja et al., 2026), InternVL3.5 (Zhu et al., 2025), Janus-Pro (Chen et al., 2025)), and one chemistry-specific model (ChemVLM (Li et al., 2025b)). All models are evaluated in the zero-shot setting with majority-vote accuracy over three independent trials, following common practice in multimodal benchmark evaluation (Yue et al., 2024). A question counts as incorrect when the three trials yield three different answers or when an output cannot be parsed into an option. We use reasoning-first prompting, instructing each model to analyze structural features and their mechanistic roles before producing a final answer. As a human reference, the eight chemistry experts answer the same 600-question subset used in Section 4.3, following prior benchmarks with expert baselines on a subset (Mirza et al., 2025). These results serve as

Model	Size	Avg.	Reasoning Type			Entry Type	
			Mech.	Design	Cond.	Mol.	Rxn.
Random baseline	–	25.0	25.0	25.0	25.0	25.0	25.0
<i>Proprietary VLMs</i>							
GPT-5.4	–	78.6	75.3	84.2	68.6	81.1	77.6
GPT-5	–	77.4	72.9	83.6	78.4	80.2	76.2
GPT-4o	–	68.1	65.6	72.5	60.8	68.6	67.9
Claude Sonnet 4.6	–	78.2	76.6	81.1	70.6	79.9	77.5
Gemini-3.1-Pro	–	78.0	76.1	80.6	78.4	80.2	77.1
<i>Open-Source VLMs</i>							
Qwen2.5-VL	7B	42.6	38.5	48.8	39.2	41.4	43.1
Qwen3-VL	8B	52.5	47.7	59.5	49.0	51.8	52.7
Qwen3-VL	30B	58.8	54.7	65.0	54.9	62.4	57.4
Phi-4-reasoning-vision	15B	51.3	50.6	52.2	51.0	45.9	53.4
InternVL3.5	8B	48.0	45.6	52.0	43.1	47.9	48.1
Janus-Pro	7B	23.2	20.0	27.5	25.5	26.3	22.0
<i>Chemistry-Specific VLMs</i>							
ChemVLM	8B	34.0	33.2	35.2	33.3	28.4	36.2

Table 3: Performance on ChemExplain (majority-vote accuracy, %, 3 trials). Mech. = Mechanistic, Cond. = Condition, Mol. = molecule, Rxn. = reaction. Best per column in **bold**. All results are on the full 1,180-question benchmark. Human expert performance is reported separately in Appendix Tab. 6.

a reference for answerability and difficulty rather than a direct human-model comparison.

Overall performance. Tab. 3 presents the main results. On the 600-question subset, the experts average 73.3% and range from 68.3% to 80.0% (Appendix Tab. 6), confirming that the questions are answerable for trained chemists but remain challenging. Differences among experts partly reflect domain specialization across materials, polymer, inorganic, and other areas of chemistry: on questions within his own domain, one expert reaches 94.1% (Appendix D.2). On the full benchmark, the top three models cluster near 78% (GPT-5.4: 78.6%, Claude Sonnet 4.6: 78.2%, Gemini-3.1-Pro: 78.0%), with no statistically significant pairwise differences detected (Appendix E.8). The strongest open-source model, Qwen3-VL-30B, reaches 58.8%, nearly 20 points lower. Scaling Qwen3-VL from 8B to 30B improves accuracy from 52.5% to 58.8%, but leaves a substantial gap. Appendix E.9 reports accuracy by reasoning type.

Domain-specific VLM remains weak on explanatory attribution. ChemVLM reaches 34.0%, below every general-purpose open-source model except Janus-Pro. Its errors span perception, knowledge, and attribution (Section 5.3), so the performance gap is not limited to the final attribution step. Model size does not fully explain the result: the size-matched InternVL3.5-8B and Qwen3-VL-8B reach 48.0% and 52.5%, respectively. ChemVLM is trained mainly for structure recognition and exam-style QA, whereas ChemExplain requires visually grounded reasoning from a structural feature to an observed effect.

Model	With image	Text-only	Δ
GPT-5	77.0	23.5	-53.5
Claude Sonnet 4.6	74.0	20.5	-53.5
Gemini-3.1-Pro	78.5	24.5	-54.0

Table 4: Text-only control on the 600-question expert subset, removing only the structure image.

5.2 Visual Necessity

Multimodal multiple-choice benchmarks can contain textual shortcuts that allow questions to be answered without the image (Yue et al., 2025). We test the visual necessity targeted by the anti-shortcut design (Section 3.3) on the 600-question expert subset (Section 4.3). We remove only the structure image and keep the question and all four options unchanged, which preserves all lexical, stylistic, and elimination cues. As shown in Tab. 4, accuracy falls by 53.5–54.0 points across the three models, from 74.0–78.5% with the image to 20.5–24.5% without it. Without the image, all three models perform below the 25% random baseline, suggesting that the question and options provide little exploitable information for these models. The substantial performance drop confirms that their success depends largely on the depicted structure.

5.3 VLM Error Analysis

To understand where in the reasoning chain (perception $\rightarrow$ knowledge $\rightarrow$ attribution) models fail, we adapt standard multimodal error taxonomies (Yue et al., 2024, 2025) into a domain-specific taxonomy. A chemistry faculty member manually analyzes responses from three representative models—GPT-5.4 (best closed-source), Qwen3-VL-8B (representative open-source), and ChemVLM (chemistry-specific)—on the same 600-question stratified subset (Section 4.3), categorizing each incorrect response by its reasoning trace as a *perception error* (the model misidentifies or overlooks visible structural features), a *knowledge gap* (it perceives the features but lacks the chemical knowledge linking them to the effect), or an *attribution error* (it perceives the relevant features and has the necessary knowledge but misjudges which feature is primarily responsible in this structural context).

Three chemistry experts independently reclassified all 850 incorrect responses. They assigned the same error type in 96.5% of cases, with Gwet’s $AC_2=0.984$ (95% CI [0.966, 1.000]). This high agreement supports the consistency of the perception, knowledge, and attribution taxonomy.

Error distribution. Across the three models, attribution accounts for 48% of all incorrect responses,

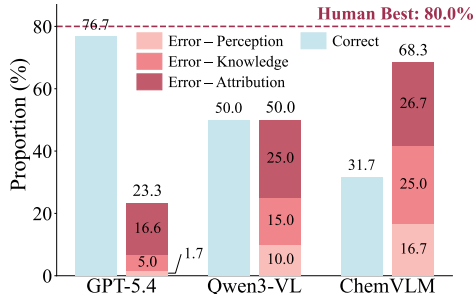


Figure 4: Accuracy and error type breakdown for three representative models on the 600-question evaluation subset. Dashed line: best human expert accuracy.

followed by knowledge gaps (32%) and perception errors (20%) (Fig. 4). The error profiles differ with model performance. GPT-5.4 makes only 10 perception errors on the subset, while 100 of its 140 errors (71.4%; 95% Wilson CI [63.5, 78.3]) are attribution errors, a proportion that remains 72.0% on questions with strong expert consensus in the blind phase (Appendix D.1). In these attribution-classified failures, the reasoning traces correctly identify relevant structural features and apply plausible chemistry, but select the wrong feature as the primary explanation. Providing the structural features and the relevant chemistry corrects most perception errors and knowledge gaps, but attribution errors largely persist (Appendix D.4). A chemistry expert answering without the source paper resolves 60 of GPT-5.4’s 100 attribution errors, and most of those he misses fall outside his domain (Appendix D.2). Qwen3-VL-8B shows a more balanced error profile, with attribution accounting for 50% of its failures. ChemVLM makes 100 perception errors and 150 knowledge errors, showing that its low performance cannot be attributed to the final attribution step alone. Appendix E.6 presents representative examples of all three error types.

5.4 Agentic Construction Pipeline Analysis

We analyze the main construction choices on 100 randomly sampled source papers (Tab. 5) and examine attrition across the full corpus (Fig. 3). Repeating the subset analysis with Claude Sonnet 4.6 supports the same design choices (Appendix E.4), and Appendix D.3 shows that current VLMs struggle to distinguish structural analogs and to identify explanatory evidence in text.

Image-first representation preserves structurally complex molecules. A central design choice is whether to retain segmented structure images or first convert them into SMILES strings.

Configuration	Structures	Ent. Match	Expl.
Full pipeline (ours)	489	76%	43
w/o image-first (SMILES mode)	154	—	2
w/o VLM identifier	489	50%	34
w/ MolCoref identifier	489	35%	31
w/o context expansion	489	76%	26

Table 5: Construction pipeline ablation on 100 randomly sampled source papers. Each row below the full pipeline removes or replaces one component, and the columns report pipeline outputs. Structures: individual molecular or reaction images retained after segmentation. Ent. Match: fraction of retrieved text referring to the target molecule or reaction. Expl.: explanatory sentences.

The SMILES-first alternative, implemented with MolScribe (Qian et al., 2023b) and DECIMER (Rajan et al., 2021) plus round-trip quality control (Appendix G), discards 68.5% of the retained structures and leaves only 2 of the 43 explanatory sentences (Tab. 5). Extraction failures are particularly common for large rings, densely substituted structures, and nonstandard drawing conventions, including structures mechanistically discussed in the literature. Retaining images therefore avoids a strong selection bias and preserves structural diversity.

Identifier-based retrieval enables precise text grounding. Removing VLM-based identifier extraction lowers entity match from 76% to 50%, and substituting MolCoref (Fan et al., 2024) lowers it further to 35% (Tab. 5). MolCoref often assigns no identifier or the identifier of a neighboring compound, which misleads downstream retrieval.

Context expansion recovers explanatory content across sentence boundaries. Without context expansion, only 26 explanatory sentences are identified in the 100 sampled papers, whereas expansion raises this number to 43 (a 65% gain). Even so, the availability of explanatory text remains the primary bottleneck: in the full corpus, only 13.1% of structures with retrieved text contain at least one explanatory sentence. To verify that this attrition reflects a scarcity of explanatory text rather than a retrieval failure, a chemistry expert audited 244 candidate structures from 50 random papers and found the same rates as the full corpus: 39.8% had retrieved text, and 13.4% of these contained an explanatory sentence (13.1% in the full corpus). The discarded cases lack a structure-to-effect explanation rather than being mis-retrieved, and most are partially explanatory, naming a feature and its effect without the mechanism (Appendix D.3).

6 Discussion

What ChemExplain measures. ChemExplain evaluates explanatory attribution: identifying which visible structural feature most directly explains an observed chemical effect. Each question isolates this decision, because all four options explain the same effect through features visible in the structure (Section 3.3), so choosing among them requires judging which feature is primarily responsible. Two results show that the benchmark measures this decision reliably. The questions require the depicted structure, as removing the image lowers accuracy to 20.5–24.5% (Section 5.2). The labeled answers reflect expert consensus, as at least five of the eight experts select them for 94.0% of questions without seeing the source paper (Section 4.3).

Implications for model development. Attribution errors decrease least as models become stronger: from ChemVLM to GPT-5.4 they fall by 38%, compared with 80–90% for perception and knowledge errors, and they persist when the structure and the relevant chemistry are supplied (Section 5.3). These findings suggest that training models to compare competing explanations within a specific structural context may improve attribution. The Effect, Feature, and Mechanism fields in ChemExplain provide a natural source of supervision for this purpose. Evidence-Grounded Attribution, which links each option to evidence in the depicted structure, already improves three open-source models by 3.1–5.3 points, with the largest gain for ChemVLM (Appendix F). Chemistry-specific VLMs thus need training data pairing structures with explanations of their effects, beyond structure recognition and exam-style QA.

7 Conclusion

We introduced ChemExplain, a literature-grounded benchmark of 1,180 questions for explanatory attribution in chemistry, constructed automatically from peer-reviewed papers and assessed by eight chemistry experts. The strongest evaluated model reaches 78.6% accuracy, and reasoning-trace analysis classifies 71.4% of its errors as attribution failures. These results identify explanatory attribution as an unresolved problem in multimodal chemical reasoning and establish ChemExplain as a practical way to measure progress. Appendix K discusses design choices and future directions.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grant No. 2435819 and the National Library of Medicine (NLM) of the National Institutes of Health (NIH) under Award No. R01LM014385.

Limitations

ChemExplain evaluates explanatory attribution in a multiple-choice format, which enables automatic and reproducible scoring. Free-form explanation generation is a complementary setting, and the source evidence retained for each question could serve as its reference. The benchmark focuses on 2D structure drawings, and extending it to other representations such as spectra, 3D conformations, or crystal structures is a direction for future versions.

Ethical Considerations

All source papers are drawn from the PMC Open Access Subset, which distributes articles under various Creative Commons licenses (CC-BY, CC-BY-NC, CC-BY-SA, and others) that permit text mining and derivative works. Our benchmark construction constitutes a derivative work that extracts and transforms content from these articles. Each benchmark item retains full provenance to its source paper, enabling verification and compliance with individual article license terms. We release ChemExplain under CC-BY-NC-SA 4.0 for research purposes. Users intending commercial applications should consult the license of each source article. To support transparency and reproducibility, we will also release the construction code and the full prompt templates upon publication.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mahmoud Amiri and Thomas Bocklitz. 2025. Chemrxivquest: A curated chemistry question-answer database extracted from chemrxiv preprints. *arXiv preprint arXiv:2505.05232*.

Jyoti Aneja, Michael Harrison, Neel Joshi, Tyler LaBonte, John Langford, and Eduardo Salinas. 2026. Phi-4-reasoning-vision-15b technical report. *arXiv preprint arXiv:2603.03975*.

Anthropic. 2025. [Claude sonnet 4.6](#).

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai et al. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. 2025. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19553–19564. IEEE.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. 2024. Scholarchemqa: Unveiling the power of language models in chemical research question answering. *arXiv preprint arXiv:2407.16931*.

Google DeepMind. 2026. [Gemini 3.1 pro](#).

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. 2025. Supergqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Vincent Fan, Yujie Qian, Alex Wang, Amber Wang, Connor W Coley, and Regina Barzilay. 2024. Openchemie: An information extraction toolkit for chemistry literature. *Journal of Chemical Information and Modeling*, 64(14):5521–5534.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in neural information processing systems*, 36:59662–59688.

- Kilem Li Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4th edition. Advanced Analytics, LLC.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, Yi Li, et al. 2024. Chemeval: a comprehensive multi-level chemical evaluation for large language models. *arXiv preprint arXiv:2409.13989*.
- Joseph James. 2026. Counting on consensus: Selecting the right inter-annotator agreement metric for nlp annotation and evaluation. *arXiv preprint arXiv:2603.06865*.
- Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.
- Siddhartha Laghuvarapu, Namkyeong Lee, Chufan Gao, and Jimeng Sun. 2025. Moltextqa: A question-answering dataset and benchmark for evaluating multimodal architectures and llms on molecular structure-text understanding. *Journal of Data-centric Machine Learning Research*.
- Shi Xuan Leong, Sergio Pablo-García, Brandon Wong, and Alán Aspuru-Guzik. 2025. Mermaid: Universal multimodal mining of chemical reactions from pdfs using vision-language models. *Matter*, 8(12).
- Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. 2025a. Beyond chemical qa: Evaluating llm’s chemical reasoning with modular chemical operations. *arXiv preprint arXiv:2505.21318*.
- Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. 2025b. Chemvbm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.
- David Y Liu, Xanthe Muston, Aditya Joshi, and Sebastian Sequoia-Grayson. 2026. Retell, reward, repeat: Reinforcement learning for narrative theory-informed story generation. *arXiv preprint arXiv:2601.17226*.
- Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. Moleculeqa: A dataset to evaluate factual accuracy in molecular comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3769–3789.
- Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034.
- Anu Pradhan, Alexandra Ortan, Apurv Verma, and Madhavan Seshadri. 2025. Llm-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. *arXiv preprint arXiv:2509.12382*.
- Yujie Qian, Jiang Guo, Zhengkai Tu, Connor W Coley, and Regina Barzilay. 2023a. Rxnscribe: a sequence generation model for reaction diagram parsing. *Journal of chemical information and modeling*, 63(13):4030–4041.
- Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. 2023b. Molscribe: robust molecular structure recognition with image-to-graph generation. *Journal of chemical information and modeling*, 63(7):1925–1934.
- Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. Decimer 1.0: Deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13(1):61.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First conference on language modeling*.
- Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. 2025. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 66(1):216–227.
- SCImago. 2024. SJR—SCImago journal & country rank. <http://www.scimagojr.com>. Accessed: 2025.
- Ayush Kumar Shah, Abhisek Dey, Leo Luo, Bryan Amador, Patrick Philippy, Ming Zhong, Siru Ouyang, David Mark Friday, David Bianchi, Nick Jackson, et al. 2025. Multimodal search in chemical documents and reactions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4030–4034.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.
- Geemi P Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. 2025. Chemlit-qa: a human evaluated dataset for chemistry rag tasks. *Machine Learning: Science and Technology*, 6(2):020601.
- Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen’s kappa and gwet’s ac1 when calculating interrater reliability coefficients: a study conducted with personality disorder samples. *BMC medical research methodology*, 13(1):61.
- Jiaqing Xie, Weida Wang, Ben Gao, Zhuo Yang, Haiyuan Wan, Shufei Zhang, Tianfan Fu, and Yuqiang Li. 2025. Qcbench: Evaluating large language models on domain-specific quantitative chemistry. *Journal of Chemical Information and Modeling*, 65(22):12268–12278.
- Ruiling Xu, Yifan Zhang, Qingyun Wang, Carl Edwards, and Heng Ji. 2025. omebench: Towards robust benchmarking of llms in organic mechanism elucidation and reasoning. *arXiv preprint arXiv:2510.07731*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. 2025. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29580–29590. IEEE.
- Zehua Zhao, Zhixian Huang, Junren Li, Siyu Lin, Junting Zhou, Fengqi Cao, Kun Zhou, Rui Ge, Tingting Long, Yuexiang Zhu, et al. 2025. Superchem: A multimodal reasoning benchmark in chemistry. *arXiv preprint arXiv:2512.01274*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in neural information processing systems*, volume 36, pages 46595–46623.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

ChemExplain: A Multimodal Benchmark for Explanatory Chemical Reasoning

Supplementary Materials

Table of Contents

- A. Extended Related Work
- B. Representative Benchmark Example
- C. Construction Pipeline Details
 - C.1 Visual Processing
 - C.2 Text Grounding
 - C.3 Question Generation
- D. Additional Analysis of VLM Capabilities
 - D.1 Human Expert Performance
 - D.2 Has GPT-5.4 Closed the Attribution Gap?
 - D.3 VLM Limitations Revealed by the Construction Framework
 - D.4 Interventional Validation of the Error Taxonomy
- E. Additional Results
 - E.1 Agreement Metric
 - E.2 Sample Balance
 - E.3 Agreement Across Judge Models
 - E.4 Ablation Replication Across Model Families
 - E.5 Rejected Question Analysis
 - E.6 VLM Error Case Studies
 - E.7 Rejected Question Examples
 - E.8 Accuracy Uncertainty Across Source Papers
 - E.9 Accuracy by Reasoning Type
- F. Evidence-Grounded Attribution
- G. SMILES-First Pipeline
- H. Prompt Templates
- I. Human Evaluation Platform
- J. Benchmark Question Examples
- K. Design Considerations and Future Work

A Extended Related Work

Benchmarks for Chemical and Scientific Reasoning. A broad literature evaluates LLMs and VLMs on chemistry, but most benchmarks target prediction or retrieval rather than explanation of observed chemical effects. This distinction matters because a model can produce a correct product or pathway without identifying which structural feature actually drives the effect (Mirza et al., 2025). Broad scientific benchmarks such as GPQA (Rein et al., 2024), SuperGPQA (Du et al., 2025), and SciBench (Wang et al., 2023) evaluate advanced reasoning across disciplines, including chemistry, but they do not involve molecular diagrams or reaction schemes. Chemistry-focused benchmarks such as ChemLLMBench (Guo et al., 2023), ChemBench (Mirza et al., 2025), and ChemEval (Huang et al., 2024) cover property prediction, structure elucidation, and name conversion (Runcie et al., 2025; Lu et al., 2024; Xie et al., 2025; Guo et al., 2024), while ChemCoTBench (Li et al., 2025a), oMeBench (Xu et al., 2025), and SUPERChem (Zhao et al., 2025) focus on reaction outcome prediction and pathway generation. Literature-grounded chemistry QA benchmarks (Laghuvarapu et al., 2025; Amiri and Bocklitz, 2025; Chen et al., 2024; Wellawatte et al., 2025) derive questions from papers, but they mainly evaluate factual retrieval. Across these benchmarks, the ability to attribute an observed chemical effect to the most responsible visible structural feature remains largely unmeasured. ChemExplain addresses this gap.

Multimodal Chemistry Information Extraction and Literature-Based Benchmarks. A related line of work studies multimodal chemistry information extraction from scientific documents. These systems have progressed from text-based entity recognition (Swain and Cole, 2016) to multimodal pipelines (Fan et al., 2024; Leong et al., 2025; Shah et al., 2025) that recover molecular structures, reaction records, and associated text from papers. However, such pipelines typically retrieve all text linked to a molecule or reaction, without distinguishing explanatory evidence from procedures,

observations, or characterization details. Existing quality control methods verify either the chemical validity of extracted structures (Shah et al., 2025) or the entity alignment between images and text (Leong et al., 2025), but they do not validate whether the retrieved text supports a structure-to-effect explanation. Literature-based chemistry benchmarks (Laghuvarapu et al., 2025; Amiri and Bocklitz, 2025; Chen et al., 2024; Wellawatte et al., 2025) often rely on LLM-generated answers, which introduces an additional risk of hallucinated rationales (Wellawatte et al., 2025). Agentic pipelines such as AutoConverter (Zhang et al., 2025) and MicroVQA (Burgess et al., 2025) convert human-written question-answer pairs into multiple-choice questions. Our framework instead validates explanatory evidence and derives each correct answer directly from the source paper.

B Representative Benchmark Example

Fig. 6 shows a complete ChemExplain question. Given a molecular structure image and an observed effect (high durability and high reaction quantum yield), the model must select the explanation that most directly accounts for the effect. Each of the four candidate explanations identifies a different structural feature visible in the depicted Re(I)-porphyrin complex and describes how it leads to the observed effect. The correct answer (C) is grounded in the source paper evidence; the three distractors cite other features genuinely present in the structure—the bromide ligand (A), the carbonyl ligands (B), and the Zn(II)-metalated N4 core (D)—with plausible but incorrect mechanisms.

C Construction Pipeline Details

This section provides the implementation details of the three construction stages described in Sections 3.1–3.3.

C.1 Visual Processing

Figure classification and segmentation. A ResNet-50 classifier (He et al., 2016) assigns each figure to one of four categories: single molecule, multiple molecules, reaction scheme, or non-chemistry. Non-chemistry figures, such as spectra, plots, and experimental images, are discarded. The remaining chemistry figures are then segmented into individual visual instances. Multi-molecule figures are processed with DECIMER Segmentation (Rajan et al., 2021), whereas reaction

schemes are decomposed into reactant, product, and condition components using RxnScribe (Qian et al., 2023a). Single-molecule figures require no segmentation.

Identifier extraction. Chemistry papers typically annotate compounds with short identifiers that recur throughout the body text. MolCoref (Fan et al., 2024) recovers them for only $\sim 27\%$ of molecule figures and $\sim 15\%$ of reaction schemes in our evaluation. Its extraction assumes regular grid-style layouts, whereas real paper figures often contain arrows, annotations, and mixed content, so many structures receive either no identifier or one belonging to a neighboring compound. We therefore extract the identifier directly from the figure with a VLM (GPT-4o-mini). The target structure is highlighted with a red bounding box in the original figure, and the VLM returns the corresponding identifier.

Structure validation. After segmentation and identifier extraction, the quality control agent verifies that each candidate region is a valid 2D chemical structure. This step removes 3D renderings, crystal packing diagrams, and schematic illustrations that may survive the upstream classifier.

C.2 Text Grounding

Identifier-based retrieval. For each structure, the text grounding agent forms a retrieval package consisting of the identifier, the structure image, the figure caption, and in-text figure references such as “as shown in Fig. 3”. An LLM (GPT-4o-mini) then searches the full paper for sentences that discuss the target structure. Matching requires boundary-aware handling of short alphanumeric identifiers. For instance, identifier “1” should match “compound 1” but not the substring “DFA1”.

Context expansion. Explanatory evidence in chemistry papers often spans adjacent sentences: one sentence states the observed effect, while a neighboring sentence provides the mechanism. For each retrieved sentence, the agent locates the containing paragraph and selects semantically related surrounding sentences with an LLM, forming an expanded evidence block. If any sentence in that block contains explanatory content, the entire block is retained, and partially explanatory sentences within it are kept as supporting context rather than as the basis for a question. Section 5.4 quantifies the gain from this step.

C.3 Question Generation

Explanatory decomposition. Given a validated pair of structure image and explanatory text, an LLM (GPT-5.1) decomposes the explanatory text into three structured fields: (1) Effect: the observed effect, described without naming any structural feature (e.g., “resists oxidative degradation”); (2) Feature: the structural feature identified in the source paper as the primary cause (e.g., “bulky tert-butyl groups at ortho/para positions”); and (3) Mechanism: how the feature leads to the effect (e.g., “steric shielding around phenolic OH slows oxidative attack”). The same call also returns three to five other structural features that are visible in the image but are not the cause given in the source text. If the explanatory text is too vague to yield all three fields, the candidate is discarded rather than completed by the model.

Question and distractor generation. The question text is generated from the Effect field only. The LLM does not receive the Feature or Mechanism fields during question generation, preventing the correct answer from leaking into the question text. The correct answer is then constructed from Feature + Mechanism. For the three distractors, the LLM draws on the other visible features identified during decomposition and pairs each with a chemically plausible mechanism for the same effect. To ensure faithfulness, the correct answer’s mechanism must trace to the source paper, and the LLM is not permitted to introduce mechanisms beyond what the paper provides. Appendix H gives the full prompt templates, including the complete set of hedging terms excluded under the equal-confidence constraint, such as “indirect” and “secondary”.

D Additional Analysis of VLM Capabilities

D.1 Human Expert Performance

Eight chemistry domain experts answer the same 600-question stratified subset used for quality validation (Section 4.3). Tab. 6 reports their accuracy by reasoning category and entry type. These results are a reference for answerability and difficulty, not a direct comparison to the model results in Tab. 3, which cover the full benchmark.

Blind-phase agreement. Because experts answer each question before they see the source evidence (Section 4.3), their answers test whether the labeled answer can be identified from the question and the depicted structure alone. At least five of the

Human Experts	Avg.	Mech.	Design	Cond.	Mol.	Rxn.
Average	73.3	71.0	75.9	72.8	72.2	74.4
Best	80.0	77.4	82.8	80.4	70.0	90.0
Worst	68.3	67.7	69.0	68.6	66.7	70.0

Table 6: Human expert accuracy (%) on the 600-question stratified subset, reported for eight chemistry domain experts. Mech.: Mechanistic; Cond.: Condition; Mol.: molecule; Rxn.: reaction.

Metric	Value
Mean expert accuracy	73.3%
Labeled answer chosen by ≥ 5 of 8 experts	94.0%
Labeled answer is the most-voted option	99.7%
Mean vote lead over the next option	4.45
Median vote lead over the next option	4

Table 7: Agreement of the eight experts’ blind answers with the labeled answer on the 600-question subset. Tab. 8 gives the full distribution of expert votes.

eight experts select the labeled answer for 94.0% of the 600 questions (Tab. 7). If a distractor were an equally valid answer, experts who miss the labeled answer would converge on it. On the 414 questions that two or three experts answer incorrectly, however, these experts select different distractors in 321 cases (77.5%). The labeled answer is therefore the consensus of chemists who have not seen the source paper, and individual errors reflect the limits of each expert’s knowledge, as Expert A’s 94.1% accuracy within his own domain illustrates (Appendix D.2).

GPT-5.4’s attribution errors occur mainly on questions that most experts answer correctly without the source paper (Tab. 8). Of its 140 errors, 125 fall on questions that at least five experts answer correctly, and attribution accounts for 72.0% of these errors, compared with 71.4% overall. These errors therefore reflect a limitation of the model rather than uncertainty in the labeled answers.

D.2 Has GPT-5.4 Closed the Attribution Gap?

On the full 1,180-question benchmark, GPT-5.4 reaches 78.6% accuracy. On the 600-question stratified subset (Section 4.3), it reaches 76.7%, exceeding the human expert average (73.3%) and nearly matching the best individual expert (80.0%). Does this mean current VLMs have reached human-level explanatory attribution? To investigate, we move beyond overall accuracy and compare the error profiles of GPT-5.4 and a human expert on this subset, categorizing each incorrect response into perception error, knowledge gap, and attribution error

Votes	Questions	GPT-5.4 errors	Attr. errors
8/8	50 (8.3%)	5	4
7/8	100 (16.7%)	15	12
6/8	210 (35.0%)	45	36
5/8	204 (34.0%)	60	38
$\leq 4/8$	36 (6.0%)	15	10
$\geq 5/8$	564 (94.0%)	125	90
Overall	600	140	100

Table 8: Distribution of blind expert votes for the labeled answer on the 600-question subset, with GPT-5.4 errors and the subset of these errors classified as attribution failures (Section 5.3). Votes: number of the eight experts selecting the labeled answer. Attr.: attribution.

following the taxonomy in Section 5.3. We select Expert A, a faculty member with 15+ years of experience specializing in organic and medicinal chemistry (71.7% overall accuracy). During evaluation, Expert A annotated each question with his confidence level and whether the topic fell outside his primary research area, allowing us to separate domain knowledge gaps from attribution errors.

The expert’s advantage: attribution depth. We examine how Expert A performs on GPT-5.4’s 140 error questions (Tab. 9). Expert A answers 71.4% correctly, resolving all knowledge gaps and perception errors, and 60% of attribution errors. Of the remaining attribution errors, 75% are on questions outside his expertise. Fig. 8 illustrates a representative case.

Error Type	Count	Expert A Correct
Attribution Error	100	60
Knowledge Gap	30	30
Perception Error	10	10
Total	140	100

Table 9: Expert A’s accuracy on the 140 questions GPT-5.4 answers incorrectly, grouped by error type (Section 5.3).

The model’s advantage: domain breadth. Conversely, of the 130 questions that GPT-5.4 answers correctly but Expert A does not, 120 (92%) fall outside Expert A’s domain, suggesting that these errors reflect unfamiliarity with specific chemistry subdomains rather than a failure in explanatory attribution. Fig. 9 shows a representative case.

Decomposing overall accuracy. The findings above suggest that comparable overall accuracy masks different capability profiles. Tab. 10 quantifies this by splitting the 600 questions by whether Expert A annotated the question as within or out-

Subset	Expert A	GPT-5.4	Δ
Within-domain	94.1	70.6	$\uparrow 23.5$
Out-of-domain	42.3	84.6	$\downarrow 42.3$
Overall	71.7	76.7	$\downarrow 5.0$

Table 10: Accuracy (%) of Expert A and GPT-5.4 on within-domain and out-of-domain subsets ($n=340$ and $n=260$ respectively), defined by Expert A’s self-annotations. Δ : Expert A accuracy minus GPT-5.4 accuracy.

side his expertise.

On questions within his expertise, Expert A achieves 94.1%, far exceeding GPT-5.4’s 70.6%. His overall accuracy (71.7%) is reduced by out-of-domain questions (42.3%), on which GPT-5.4 maintains 84.6%. The answer to the opening question is clear: GPT-5.4 has not closed the attribution gap. Its competitive overall accuracy reflects broad domain coverage, not stronger explanatory attribution. The key direction for improving VLM performance on ChemExplain is not broader domain knowledge, but stronger explanatory attribution, the ability to determine which structural feature is most directly responsible for an observed effect.

D.3 VLM Limitations Revealed by the Construction Framework

The multi-agent framework relies on VLMs at multiple stages of the construction process, from aligning structures to text to evaluating explanatory quality. Analyzing where VLMs fail during construction reveals specific capability gaps of current models.

Figure-text alignment. Chemistry papers often contain structural analogs that differ by a single substituent or functional group. Aligning each structure in a figure to its discussion in the text requires distinguishing these variants. Without identifier-based retrieval, the VLM achieves only 50% entity match (Tab. 5): half of the retrieved text refers to a different compound rather than the target molecule or reaction. This indicates that current VLMs cannot reliably distinguish structural analogs from images and align them to the corresponding text in the paper.

Distinguishing explanatory from descriptive text. In chemistry papers, most text that mentions a molecule or reaction is procedural or descriptive rather than explanatory (Section 3.2). Identifying explanatory sentences is the largest bottleneck in the construction framework: of 9,670 structures

with retrieved text, only 1,264 (13.1%) contain at least one explanatory sentence (Section 4.1). Most discarded cases are partially explanatory: they identify a relevant structural feature and describe an observed effect, but do not provide the connecting mechanism, as in “the bulky substituent improves selectivity” without explaining how steric shielding leads to this outcome. Other discarded cases discuss a mechanism without linking it to any visible feature, as in PMC9719906 (chloroquine), where the retrieved text describes antiviral mechanisms only. This task is also difficult for current VLMs when explanatory content spans adjacent sentences, for example when the effect and the mechanism appear in different sentences of the same paragraph. On the 100-paper subset, 40% of explanatory sentences are missed when sentences are evaluated in isolation rather than in context, which context expansion recovers (Tab. 5). This limitation is consistent with the evaluation finding in Section 5.3: on the benchmark, VLMs struggle most with explanatory attribution, determining which structural feature is most directly responsible for an observed effect. In the construction framework, VLMs similarly struggle with identifying explanatory evidence. Improving VLM capabilities in explanatory reasoning would therefore benefit not only benchmark performance but also the ability to extract explanatory evidence from scientific literature.

D.4 Interventional Validation of the Error Taxonomy

Section 5.3 assigns error types from reasoning traces. To test whether these types reflect where models fail, we add expert-verified oracle information to the prompt. If the error types reflect where models fail, structural descriptions should correct perception errors and chemical knowledge should correct knowledge gaps, whereas attribution errors, whose traces already show correct perception and knowledge, should persist when both are supplied. We evaluate GPT-5.4 and Qwen3-VL-8B on the questions they answer incorrectly in the 600-question subset (140 and 300, of which 100 are answered incorrectly by both), together with 100 control questions answered correctly by both models, for 440 distinct questions. Each question is evaluated three times under four conditions and scored by majority vote: a rerun of the original prompt (C0), descriptions of the structural features referenced by the options (C1), chemical knowl-

edge linking each option’s feature to the effect (C2), and both (C3). Two-sided exact McNemar tests compare each condition with C0.

Oracle construction and verification. Claude Sonnet 4.6 drafts one C1 structural description for each of the 440 questions from the molecular or reaction image and the four options, without access to the labeled answer. A fixed template requires one description of each structural feature referenced by an option and prohibits comparative terms such as “stronger,” “dominant,” or “more important.” A chemistry expert blind to the labeled answer reviews all 440 descriptions for structural correctness, localization, completeness, and option neutrality, accepting 421 (95.7%) as written and revising 19 (4.3%). A second expert independently reviews all 440 descriptions and revises three more. Both experts judge every description to be neutral across options. For C2, each option is paired with a short, expert-verified statement of the chemical knowledge linking its feature to the effect, including the conditions under which the underlying principle applies. These statements are matched in level of detail across options and do not rank the options, leaving the judgment of which feature is primarily responsible to the model.

Results. The oracle information corrects perception errors and knowledge gaps but leaves attribution errors largely uncorrected (Tab. 11). Structural descriptions raise the correction rate of perception errors, the percentage answered correctly after the intervention, from 20.0% under C0 to 80.0% for GPT-5.4 and from 15.0% to 70.0% for Qwen3-VL-8B. Chemical knowledge raises that of knowledge gaps from 16.7% to 70.0% and from 13.3% to 65.6%. Both effects are significant, and neither type of information significantly affects the other error types (Tab. 12). For attribution errors, supplying both raises the correction rate from 14.0% to 19.0% and from 12.0% to 16.0%, and no condition has a significant effect. Accuracy on the control questions stays within one point of C0 under every condition (Tab. 13), so the oracle information does not alter answers that are already correct. The error types assigned from reasoning traces therefore reflect where the models fail, and attribution errors arise from misjudging which feature is primarily responsible rather than from undetected failures of perception or knowledge.

Model	Error type	N	C0	C1: Structure	C2: Knowledge	C3: Both
GPT-5.4	Perception	10	20.0 (2/10)	80.0 (8/10)	20.0 (2/10)	90.0 (9/10)
	Knowledge	30	16.7 (5/30)	23.3 (7/30)	70.0 (21/30)	76.7 (23/30)
	Attribution	100	14.0 (14/100)	17.0 (17/100)	16.0 (16/100)	19.0 (19/100)
Qwen3-VL-8B	Perception	60	15.0 (9/60)	70.0 (42/60)	16.7 (10/60)	76.7 (46/60)
	Knowledge	90	13.3 (12/90)	17.8 (16/90)	65.6 (59/90)	71.1 (64/90)
	Attribution	150	12.0 (18/150)	14.7 (22/150)	14.0 (21/150)	16.0 (24/150)

Table 11: Correction rates (%) of originally incorrect questions under each condition, with counts in parentheses. C0 reruns the original prompt, C1 adds structural descriptions, C2 adds chemical knowledge, and C3 adds both. Bold marks a significant difference from C0 (two-sided exact McNemar test, $p < 0.05$).

Model	Error type	C1 vs. C0	C2 vs. C0	C3 vs. C0
GPT-5.4	Perception	$b = 6, c = 0, p = .031$	$b = 1, c = 1, p = 1.000$	$b = 7, c = 0, p = .016$
	Knowledge	$b = 4, c = 2, p = .688$	$b = 18, c = 2, p < .001$	$b = 20, c = 2, p < .001$
	Attribution	$b = 8, c = 5, p = .581$	$b = 7, c = 5, p = .774$	$b = 10, c = 5, p = .302$
Qwen3-VL-8B	Perception	$b = 35, c = 2, p < .001$	$b = 5, c = 4, p = 1.000$	$b = 39, c = 2, p < .001$
	Knowledge	$b = 8, c = 4, p = .388$	$b = 49, c = 2, p < .001$	$b = 54, c = 2, p < .001$
	Attribution	$b = 9, c = 5, p = .424$	$b = 8, c = 5, p = .581$	$b = 11, c = 5, p = .210$

Table 12: Two-sided exact McNemar tests comparing each condition with C0. b : questions correct under the condition but incorrect under C0. c : questions correct under C0 but incorrect under the condition.

Model	C0	C1	C2	C3
GPT-5.4	96.0	97.0	96.0	97.0
Qwen3-VL-8B	92.0	93.0	92.0	93.0

Table 13: Accuracy (%) on 100 control questions that both models originally answer correctly. No oracle condition differs significantly from C0 (two-sided exact McNemar tests, all $p > 0.05$).

Dimension	Benchmark	Subset	Δ	p
Distractor Quality	4.447	4.450	+0.003	0.720
Meaningfulness	4.447	4.470	+0.023	0.400
Answer Unambiguity	4.480	4.460	-0.020	0.440
Struct. Necessity	4.505	4.630	+0.125	0.021

Table 14: Judge scores on the 1–5 Likert scale for the full benchmark and the 600-question expert subset. Δ : Subset minus Benchmark. p : two-sided Mann–Whitney U test.

E Additional Results

E.1 Agreement Metric

We report Gwet’s AC_2 with quadratic weights over the full category set $[1, 2, 3, 4, 5]$ (Gwet, 2014). Quality ratings in Tab. 2 cluster in the upper part of the Likert scale, which triggers the prevalence paradox (Wongpakaran et al., 2013; Feinstein and Cicchetti, 1990): when raters agree on most items and those items concentrate in one category, chance-corrected coefficients such as weighted κ and Krippendorff’s α subtract an inflated chance term and can report low agreement despite near-identical ratings. AC_2 estimates the chance term from the observed rating distribution and remains stable in this regime. Recent NLP evaluation studies likewise find AC_2 stable under skewed rating distributions (Pradhan et al., 2025; James, 2026; Liu et al., 2026). Confidence intervals follow from the analytic standard errors in Tab. 2; the interval for the average uses the mean of the four per-dimension standard errors. The largest standard

error for human-LLM AC_2 is 0.022, corresponding to an approximate 95% confidence-interval half-width of 0.043. The judge’s scores agree with blind expert scores on rejected questions as well as accepted ones (Appendix E.5).

E.2 Sample Balance

Tab. 14 compares the judge’s scores for the 600-question subset against the full benchmark. Holm–Bonferroni correction over the four comparisons compares the smallest p -value against $0.05/4 = 0.0125$ and stops if it does not pass. The smallest value is 0.021, so none of the four comparisons is statistically significant at the 5% level.

E.3 Agreement Across Judge Models

To assess whether the screening depends on the choice of judge, the 600-question expert-rated subset is scored by three judges from different model families, GPT-5.2, Claude-Opus-4-7, and Gemini-3.1-Pro, each following the protocol of Section 4.3.

Judge (H-L AC ₂)	Distr.	Mean.	Unamb.	Struct.	Avg.
GPT-5.2 (orig.)	0.948	0.936	0.929	0.861	0.919
Claude-Opus-4-7	0.933	0.939	0.926	0.952	0.938
Gemini-3.1-Pro	0.952	0.940	0.932	0.973	0.949

Table 15: Human-LLM agreement (Gwet’s AC₂) on the 600-question expert subset for three judges from different model families. Distr.: Distractor Quality; Mean.: Meaningfulness; Unamb.: Answer Unambiguity; Struct.: Structure Necessity.

Config (Claude-Sonnet-4-6)	Struct.	Ent. Match	Expl.
Full pipeline	489	78%	41
w/o image-first (SMILES)	154	—	3
w/o VLM identifier	489	45%	31
w/ MolCoref identifier	489	36%	27
w/o context expansion	489	79%	24

Table 16: Ablation on the same 100-paper subset as Tab. 5, repeated with Claude-Sonnet-4-6. Struct.: individual molecular or reaction images retained after classification and segmentation. Ent. Match: fraction of retrieved text correctly referring to the target molecule or reaction. Expl.: explanatory sentences that qualify for question generation.

Tab. 15 reports human-LLM agreement (Gwet’s AC₂) for each judge. All three reach comparably high agreement with the human experts (average AC₂ of 0.919, 0.938, and 0.949, respectively), indicating that the agreement reflects the quality criteria rather than the choice of judge.

E.4 Ablation Replication Across Model Families

The ablation is repeated on the same 100-paper subset with Claude-Sonnet-4-6 at every stage, under the metrics and configurations of Section 5.4. Tab. 16 reports the result. Absolute counts shift with the model, but removing image-first construction, replacing VLM-based identifier extraction, and removing context expansion each degrade the corresponding metric, reproducing the pattern in Tab. 5.

E.5 Rejected Question Analysis

The LLM judge rejects 84 candidate questions during post-generation screening (Section 4.3), 83 with average scores between 3.0 and 4.0 and one below 3.0. The eight domain experts evaluate 60 of them on the four quality dimensions, blind to the judge’s decisions. Tab. 17 reports the scores. The mean expert score decreases from 4.53 on accepted questions to 3.68 on rejected questions, and the

Dimension	$\bar{H}$	$\bar{L}$	Cause (%)
Structure Necessity	2.70	2.69	76.7
Answer Unambiguity	3.89	3.93	20.0
Distractor Quality	3.90	3.85	3.3
Meaningfulness	4.23	4.23	0.0
Overall	3.68	3.68	—

Table 17: Quality scores for 60 rejected questions ($n=60$ from 84 rejected candidates, 8 human experts). $\bar{H}$: mean human expert score (1–5); $\bar{L}$: mean LLM judge score. Cause (%): dimension with the lowest LLM score, indicating the primary reason for rejection.

mean judge score decreases correspondingly from 4.50 to 3.68. The judge therefore distinguishes accepted from rejected candidates consistently with blind expert assessment.

Rejection causes. We attribute each rejection to the dimension with the lowest judge score, and Tab. 17 reports the share of rejections attributed to each dimension. Structure Necessity accounts for 76.7% of rejections: these questions receive a mean expert score of 2.70 on this dimension, compared with 4.75 for accepted questions, because they can be answered from the option text or general chemistry without the depicted structure (Fig. 13). Correct Answer Unambiguity accounts for another 20.0%, typically because another option offers an equally defensible explanation (Fig. 14). Distractor Quality and Question Meaningfulness are rarely the primary cause (3.3% and 0.0%), indicating that the generation constraints already address them.

E.6 VLM Error Case Studies

We present one representative example of each error type from the manual analysis in Section 5.3. Fig. 10 illustrates an attribution error, Fig. 11 a perception error, and Fig. 12 a knowledge gap.

E.7 Rejected Question Examples

We present representative examples of questions rejected during post-generation screening (Section 4.3), illustrating the most common rejection causes identified in Appendix E.5. Fig. 13 shows a structure necessity rejection and Fig. 14 shows an answer unambiguity rejection.

E.8 Accuracy Uncertainty Across Source Papers

We examine two properties of the accuracy estimates: whether differences between models exceed sampling variation, and whether the estimates

Model	Accuracy	Paper-level 95% CI	One question per paper
GPT-5.4	78.6	[75.7, 81.4]	78.4 [76.0, 80.9]
GPT-5	77.4	[74.4, 80.2]	77.2 [74.9, 79.6]
GPT-4o	68.1	[65.0, 71.2]	67.8 [65.3, 70.2]
Claude Sonnet 4.6	78.2	[75.4, 80.9]	78.1 [75.7, 80.7]
Gemini-3.1-Pro	78.0	[75.1, 80.7]	77.9 [75.5, 80.4]
Qwen2.5-VL 7B	42.6	[39.0, 46.2]	42.5 [39.9, 45.2]
Qwen3-VL 8B	52.5	[48.9, 56.0]	52.3 [49.6, 55.1]
Qwen3-VL 30B	58.8	[55.3, 62.2]	58.7 [55.9, 61.6]
Phi-4-reasoning-vision 15B	51.3	[47.8, 54.8]	51.0 [48.3, 53.8]
InternVL3.5 8B	48.0	[44.5, 51.4]	47.8 [45.2, 50.4]
Janus-Pro 7B	23.2	[20.3, 26.3]	23.3 [20.9, 25.6]
ChemVLM 8B	34.0	[30.7, 37.3]	33.8 [31.3, 36.3]

Table 18: Accuracy (%) on the full benchmark with paper-level uncertainty. Paper-level 95% CI: percentile interval from 10,000 bootstrap resamples of source papers. One question per paper: mean accuracy and 95% range over 1,000 resamples that keep one randomly chosen question from each paper.

depend on papers that contribute more questions. Questions from the same source paper tend to share a chemical domain, drawing style, and terminology and are therefore not independent, so we resample papers rather than individual questions. Tab. 18 reports accuracy with 95% confidence intervals from 10,000 paper-level bootstrap resamples, together with mean accuracy over 1,000 resamples that keep one randomly chosen question per paper.

Under paper-level resampling, GPT-5.4, Claude Sonnet 4.6, and Gemini-3.1-Pro are statistically indistinguishable: the 95% confidence intervals of all their paired differences include zero (GPT-5.4 vs. Claude Sonnet 4.6: +0.4 [−2.5, 3.3]; GPT-5.4 vs. Gemini-3.1-Pro: +0.6 [−2.3, 3.6]; Claude Sonnet 4.6 vs. Gemini-3.1-Pro: +0.2 [−2.4, 2.9]). Consistent with this, each of the four leading proprietary models ranks first in at least 13% of the one-question-per-paper resamples (GPT-5.4: 39.6%, Claude Sonnet 4.6: 25.8%, Gemini-3.1-Pro: 21.5%, GPT-5: 13.1%).

The estimates also do not depend on how many questions each paper contributes. Each of the 383 papers contributes 3.08 questions on average (median 3, maximum 8), and keeping one question per paper changes accuracy by at most 0.3 points for every model.

E.9 Accuracy by Reasoning Type

Design questions are the easiest category for all evaluated models and for the human experts. Mechanistic and Condition questions are more challenging, for example for Qwen3-VL-30B (54.7% and 54.9% vs. 65.0% on Design; Tab. 3) and for the human experts (71.0% on Mechanistic vs. 75.9%

on Design; Tab. 6).

F Evidence-Grounded Attribution

The error patterns in Section 5.3 suggest two directions for model development. Some perception and knowledge errors could be reduced through tool-augmented inference, for example by integrating structure recognition models or chemical knowledge bases. Attribution errors cannot be resolved by structure recognition or knowledge retrieval alone, because the difficulty is not access to the structure or the chemistry but ranking competing explanations within a specific structural context. This makes attribution a suitable target for structured supervision: the decomposition used in ChemExplain (Effect, Feature, and Mechanism) exposes intermediate reasoning stages that could support hierarchical training or reward design. Such supervision would encourage models to first identify relevant structural evidence, then establish the mechanism, and only then decide which feature is primary. This section tests whether the same ordering can improve inference through prompting alone, without fine-tuning or external tools.

Motivation and error pattern. The strongest open-source model trails the leading proprietary models by nearly 20 percentage points on the full benchmark (Tab. 3), leaving substantial room for improvement. The error analysis further classifies most of GPT-5.4’s errors as attribution failures, in which the reasoning trace identifies the relevant structural features and applies the necessary chemical knowledge but selects the wrong feature as the primary explanation. To better understand the pattern, a chemistry expert examined the reason-

ing traces for the 100 attribution errors made by GPT-5.4 on the 600-question subset. In many cases, the model favored an explanation that is broadly common in chemistry, such as electron withdrawal by a carbonyl, over a less typical feature supported by the depicted structure, such as a nonclassical C–H···O contact. This pattern suggests that generic chemical priors can override evidence specific to the structure under examination. We therefore use open-source models as a focused and reproducible test bed for evaluating whether this benchmark-derived insight can improve performance where the gap is largest.

Method. Evidence-Grounded Attribution (EGA) requires the model to ground every candidate explanation in the depicted structure before selecting an answer. For each option, the model first identifies the visual evidence that would support the named feature and then explains how that feature could produce the observed effect. It is instructed to discount options that are chemically plausible in general but are not well supported by the specific structure. After comparing the four grounded explanations, the model selects the feature-mechanism relation that best accounts for the effect. EGA uses a single model call and requires no external tools, SMILES conversion, or fine-tuning.

Results. We apply EGA to Qwen2.5-VL, Qwen3-VL-8B, and ChemVLM, covering two generations of a widely used open-source model family and one chemistry-specific model. All models are evaluated on the full benchmark using the same zero-shot, three-trial majority-vote protocol as the main experiments. EGA improves all three models by 3.1–5.3 percentage points (Tab. 19), showing that the benchmark analysis yields an actionable inference-time strategy rather than only a descriptive error taxonomy. The gains remain moderate because EGA targets attribution, while the error analysis also identifies substantial perception and knowledge failures in these models. The strongest EGA result reaches 55.6%, leaving considerable room for further improvement.

G SMILES-First Pipeline

As described in Section 3.1, the image-first pipeline retains all structures as images rather than converting them to SMILES strings. This section describes the SMILES-first alternative that was evaluated and abandoned due to the systematic loss of complex structures reported in Section 5.4.

Model	Baseline	EGA	Δ
Qwen2.5-VL (7B)	42.6	46.8	+4.2
Qwen3-VL (8B)	52.5	55.6	+3.1
ChemVLM (8B)	34.0	39.3	+5.3

Table 19: Accuracy (%) with Evidence-Grounded Attribution on the full benchmark under the protocol of Tab. 3. Δ denotes the absolute percentage-point improvement.

In this alternative, segmented structures are converted to SMILES strings before text grounding. For single molecules, MolScribe (Qian et al., 2023b) converts the segmented image to a SMILES string. For multi-molecule figures, DECIMER Segmentation (Rajan et al., 2021) first segments the figure into individual molecules, each of which is converted to a SMILES string by MolScribe. For reaction schemes, RxnScribe (Qian et al., 2023a) segments the reaction into individual components and extracts SMILES for each molecular component. If SMILES extraction fails for any component, the entire reaction is discarded.

We apply round-trip quality control to verify extracted SMILES strings. The extracted SMILES is redrawn into a standardized structure image using RDKit. DECIMER then processes this redrawn image to produce a second SMILES string. Both the original and regenerated SMILES are converted to molecular graphs using NetworkX and compared via graph isomorphism. If the similarity is at least 0.85, the structure passes validation; otherwise it is discarded. For reactions, all molecular components must individually pass this validation. Section 5.4 reports the quantitative comparison with the image-first pipeline. Fig. 7 illustrates the SMILES extraction and validation procedure together with two representative failure modes.

H Prompt Templates

This section provides the key prompt templates used in question generation (Section 3.3) and LLM-as-a-Judge screening (Section 4.3). Fig. 15 shows the Stage 1 explanatory decomposition prompt, Fig. 16 shows the Stage 2 question and distractor generation prompt, and Fig. 17 shows the LLM-as-a-Judge screening prompt.

I Human Evaluation Platform

Fig. 18 shows the blind answer phase, in which evaluators answer the question without access to the source paper text. Fig. 19 shows the evidence

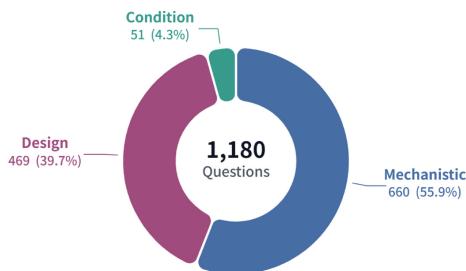


Figure 5: Composition of the 1,180-question benchmark sourced from 383 papers, distributed across three reasoning categories: Mechanistic (55.9%), Design (39.7%), and Condition (4.3%).

review and scoring phase, in which evaluators review the paper evidence and score the four quality dimensions on a 1–5 Likert scale.

J Benchmark Question Examples

Fig. 5 shows how the 1,180 questions are distributed across the three reasoning categories. We present representative ChemExplain questions across reasoning categories (Mechanistic, Design, Condition) and entry types (molecule, reaction). Each example shows the molecular or reaction image, the question, four candidate explanations, and the correct answer. Figs. 20–24 show molecule and reaction examples for each reasoning category.

K Design Considerations and Future Work

Robustness to pretraining exposure. Because the source articles are published papers, some may appear in the pretraining data of the evaluated models. Several design choices prevent a model from exploiting memorized text: the questions and distractors are newly written rather than copied from the papers, and answer positions are randomized (Section 3.3). The text-only control in Section 5.2 further shows that source-text or option recall alone cannot explain the strong image-based performance. To test whether recognizing a source paper or figure raises accuracy, we compare GPT-4o on two category-matched sets of 200 questions, drawn from papers published before and after its documented knowledge cutoff (October 1, 2023). Its accuracy is 68.5% and 68.0%, respectively, a difference of one question (95% CI [−8.6, 9.6] points), so prior exposure to the source papers does not measurably raise its accuracy. The automated pipeline can also generate future benchmark versions from papers published after a given model’s

training cutoff.

Safeguards in LLM-based construction. The framework uses LLMs and VLMs for identifier extraction, text grounding, explanatory-value assessment, and question generation, and each of these steps is constrained by source evidence or checked by quality control. The correct feature-mechanism relation must be supported by evidence from the source paper, and the question-generation model cannot introduce a mechanism absent from that evidence (Section 3.3). The anti-shortcut constraints control stylistic artifacts in the answer options, and quality control filters outputs before they advance to the next stage. Expert evaluation supports this design: accepted questions receive an average score of 4.53 out of 5 (Tab. 2), and every question in the expert subset passes the 4.0 acceptance threshold under expert scores. Future versions could further automate the explanatory-value assessment with a classifier trained on the accumulated labels.

Model-agnostic construction. The construction is designed to keep the benchmark independent of the model used for question generation. Each correct feature-mechanism relation is taken from the source paper, and the generator (GPT-5.1) only rewrites this evidence into a question and options (Section 3.3). The evaluation results confirm this property across model families. GPT-5.4, Claude Sonnet 4.6, and Gemini-3.1-Pro lie within 0.6 points of one another, Claude Sonnet 4.6 and Gemini-3.1-Pro both outperform GPT-5, and Claude Sonnet 4.6 achieves the highest Mechanistic accuracy (Tab. 3). Without the image, the three frontier models reach similar accuracies of 20.5–24.5% (Tab. 4), so the option wording is equally informative across model families. Judges from three model families also reach comparable agreement with human experts (Appendix E.3), and the construction ablations yield the same ordering of design choices with Claude Sonnet 4.6 (Appendix E.4).

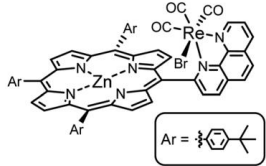
Determining the most direct explanation. An observed chemical effect may involve several structural features, so each question asks which feature is primarily responsible. The correct feature-mechanism relation is grounded in the source paper, and distractors are drawn from other features visible in the same structure that the paper does not identify as primary (Section 3.3). Experts who answer without the source paper select the labeled answer by majority on 94.0% of the expert subset (Appendix D.1), so in most questions the expla-

nation that the source paper identifies as primary is also the one that independent chemists identify. The Correct Answer Unambiguity check removes questions for which another option is equally defensible, accounting for 20.0% of rejected questions (Tab. 17). Accepted questions receive an average expert score of 4.43 out of 5 on this dimension (Tab. 2), and none of the 4,800 expert ratings falls at or below 3, the rubric level for genuine scientific debate about the primary cause (Fig. 17). The LLM judge applies the same check to every candidate question and agrees closely with the experts on this dimension ($AC_2=0.929$), so the expert results extend to the screening of the full benchmark. Future versions could also retain the removed questions through graded or multi-answer evaluation, extending coverage to effects that several features explain jointly.

Category composition. The category proportions follow the source literature. Condition questions account for 4.3% of the benchmark (51 questions) because source papers rarely state explicitly why a structural feature requires particular reaction conditions. The construction framework can increase this share by targeting papers that discuss stoichiometry, selectivity, catalyst choice, and other condition-dependent effects.

Source corpus. The source corpus is drawn from the PubMed Central Open Access Subset and restricted to journals with $SJR \geq 2.0$ (Section 4.1). The Open Access Subset provides articles under licenses that permit text mining and derivative use, and the SJR threshold favors papers that discuss structure-to-effect relationships explicitly. The resulting corpus spans organic synthesis, materials chemistry, catalysis, and medicinal chemistry. The construction pipeline can be applied to additional repositories and chemistry subdomains while preserving the license of each source paper, evidence grounding, and quality control.

Question:
The structure shown displays high durability (suppressed framework degradation) together with a high reaction quantum yield. Which structural feature most directly explains this combined effect?



Literature-grounded Evidence:
The covalent linkage between the porphyrin and the Re(CO)₃ unit through the bridging bipyridyl ligand enables rapid intramolecular electron transfer, suppressing accumulation of two-electron reduced porphyrin that causes framework hydrogenation, thereby achieving high durability and high quantum yield.

Choices:

- A. The coordinated bromide ligand on Re(I) modulates the metal-centered reduction potentials and suppresses formation of highly reactive unsaturated Re species, reducing pathways that generate radicals near the macrocycle and thereby limiting porphyrin hydrogenation while sustaining a high reaction quantum yield.
- B. The fac-Re(I)(CO)₃ fragment provides a robust low-spin d⁶ center with strong-field carbonyl ligands that stabilizes reactive reduced intermediates and channels electrons into catalytic CO₂ reduction rather than substrate-free photochemistry, producing high durability and a high reaction quantum yield.
- ✓ C. The covalent linkage placing the porphyrin photosensitizer in close proximity to the Re(I) carbonyl CO₂-reduction complex enables rapid intramolecular electron transfer after porphyrin reduction, preventing accumulation of highly reduced porphyrin that drives hydrogenation and increasing quantum yield.
- D. The Zn(II)-metallated porphyrin N₄ core tunes excited-state redox properties and stabilizes the macrocycle toward photochemical side reactions, decreasing formation of hydrogenated porphyrin products and increasing productive photoredox turnover to give high durability and quantum yield.

■ Structural Feature ■ Observed Effect (identical across all options)
■ Mechanism

Figure 6: A complete annotated ChemExplain question: a molecular structure image, an observed effect (green, identical across all options), and four candidate explanations each identifying a different structural feature (blue) and mechanism (orange). The correct answer (C, checkmark) is grounded in the source paper evidence (dashed box). Distractors reference other structural features genuinely present in the molecule.

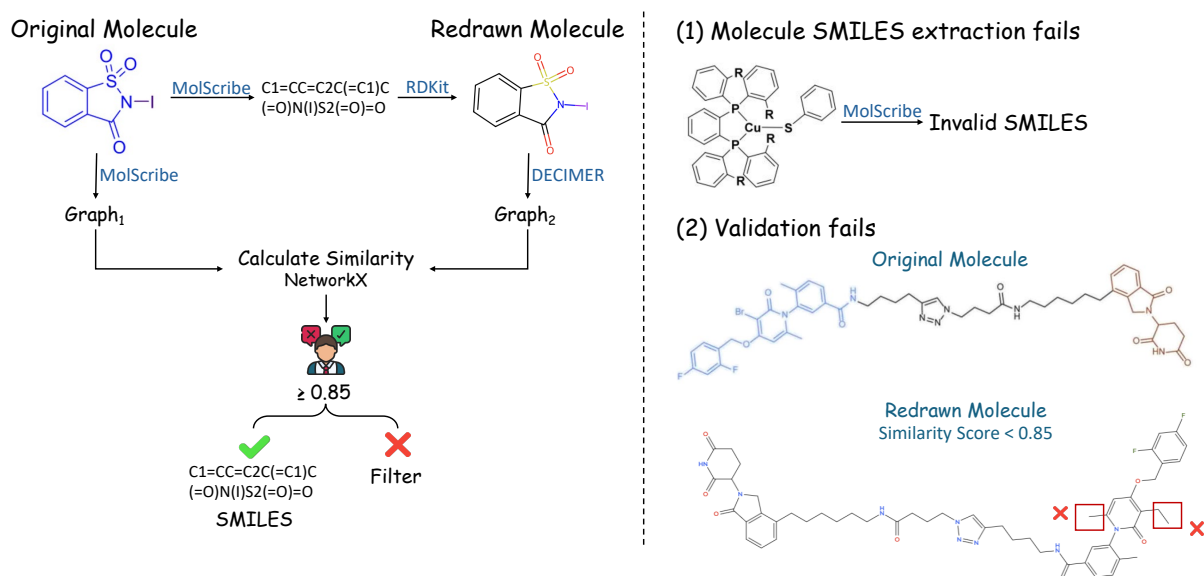


Figure 7: SMILES extraction and round-trip validation in the SMILES-first pipeline (left), with two representative failure modes (right): (1) MolScribe fails to extract a valid SMILES from a structurally complex molecule; (2) extraction succeeds but the regenerated structure differs from the original, failing the similarity check.

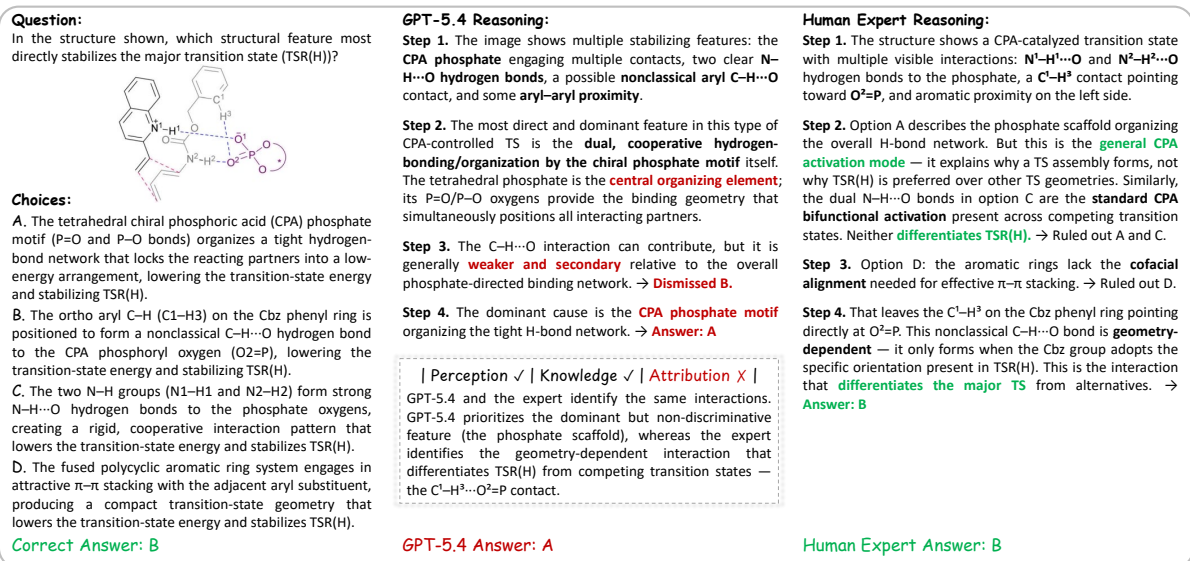


Figure 8: Representative case where GPT-5.4 commits an attribution error while Expert A, answering without the source paper, correctly identifies the responsible structural feature within his domain.

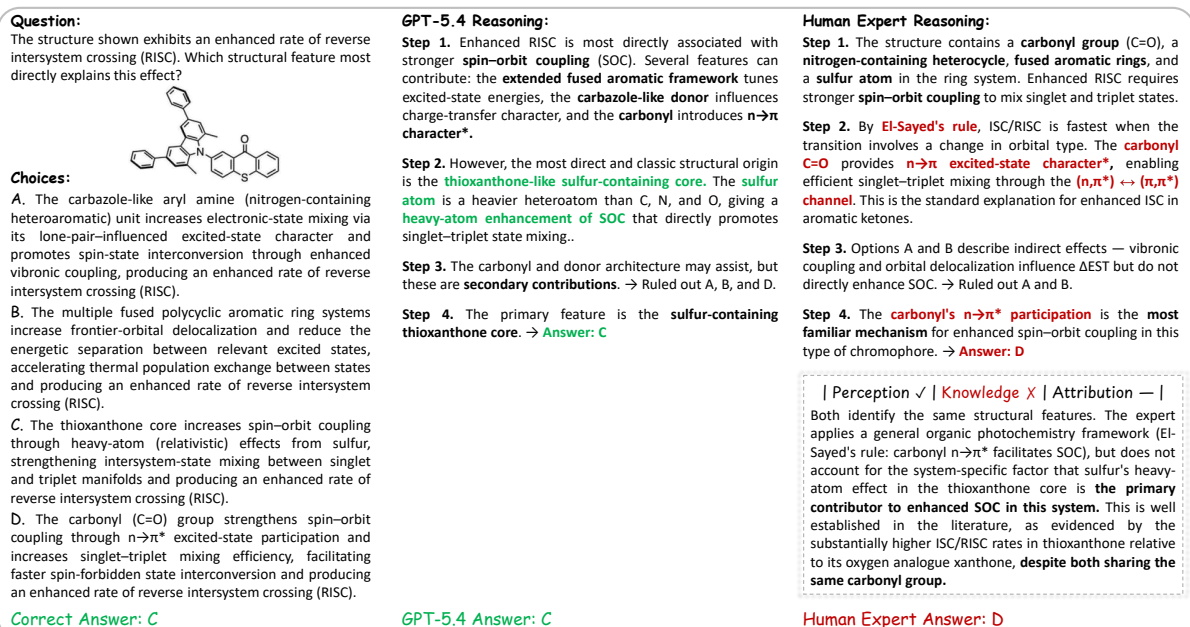
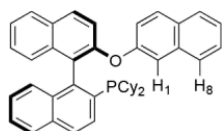


Figure 9: Representative case where GPT-5.4 answers correctly while Expert A fails due to a domain knowledge gap on a question outside his primary expertise.

Question:

In the Heck-type asymmetric arylation, this molecule enhances enantioselectivity by stabilizing the reductive-elimination transition state pathway leading to the major enantiomer. Which structural feature most directly explains this effect?

**Choices:**

- A. The PCy₂-substituted phosphine center increases steric differentiation of the diastereomeric reductive-elimination transition states through strong ligand-substrate repulsion control → enhanced enantioselectivity.
- B. The H₁ and H₈ aryl C-H groups on the ligand's naphthalene ring form two simultaneous C-H...O hydrogen-bonding interactions to the substrate carbonyl oxygen, stabilizing the major-enantiomer pathway.
- C. The extended polycyclic arene scaffold stabilizes the reductive-elimination transition state by strengthening dispersive π -surface contacts with the substrate → enhanced enantioselectivity.
- D. The aryl-O-aryl ether linkage enforces a rigid, chiral pocket that fixes the relative orientation of the biaryl fragments, creating asymmetric substrate approach geometries → enhanced enantioselectivity.

GPT-5.4 Answer: C Correct Answer: B

GPT-5.4 Reasoning:

1. Steric differentiation from the PCy₂ group certainly matters, but this describes a more generic steric repulsion model. The question specifically emphasizes stabilization, which points to attractive interactions rather than repulsion control. → Ruled out.
2. Two simultaneous C-H...O hydrogen bonds from naphthalene C-H's are much less convincing as the primary effect; such interactions are weak and geometrically specific — this is not the most characteristic explanation for these ligands.
3. C is the strongest explanation. The extended polycyclic aromatic framework provides substantial dispersion/ π -surface interactions with the substrate, selectively stabilizing one diastereomeric TS over the other.
- 4 The ether linkage helps define the ligand architecture, but it is more indirect; it organizes the scaffold rather than being the main source of TS stabilization. → Ruled out.

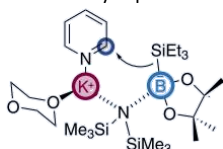
| Perception ✓ | Knowledge ✓ | Attribution ✗ |

The model correctly identifies the C-H...O mechanism in B but dismisses it as "weak" (Step 2). However, "weak" does not mean "not primary": the source paper's DFT calculations show this specific interaction directly stabilizes the major-enantiomer transition state, making it the most direct cause — not the general dispersive π -contacts the model preferred.

Figure 10: Representative attribution error from GPT-5.4 on ChemExplain. The model identifies the chemically relevant interaction and reasons plausibly about the options, yet selects the wrong feature as the primary explanation of the observed effect.

Question:

In DME, regioselective silylation occurs at the aromatic C-4 position rather than C-2; which structural feature in the structure shown most directly explains this selectivity?

**Choices:**

- A. Multiple oxygen atoms in the structure coordinate K⁺ and polarize the contact ion pair, altering aggregation and the effective nucleophilicity of the silyl anion so that attack occurs along the lowest-barrier pathway, producing silylation at C-4 over C-2.
- B. The bulky silyl substituents (SiEt₃ and SiMe₃ groups) increase steric congestion near one face of the aromatic framework, biasing approach trajectories of the silyl nucleophile and producing preferential silylation at the less hindered C-4 position.
- C. The cyclic diolato boronate (B–O₂ motif) withdraws electron density through inductive effects and organizes nearby substituents, shifting electrophilicity within the aromatic ring so that nucleophilic silyl transfer occurs preferentially at the C-4 position.
- D. Coordination of the azine nitrogen to K⁺ with solvent-separated ion pairing enforced by two bidentate DME ligands activates the azine toward nucleophilic addition and shields the C-2 site, directing bulky silyl transfer to C-4.

GPT-5.4 Answer: A **Correct Answer: D**

GPT-5.4 Reasoning:

Step 1. The key feature is the **potassium cation tightly coordinated by multiple heteroatoms**: the **azine N** and several **O atoms** from the chelating alkoxide/boronate environment. In DME, this affects the **ion-pairing/aggregation state** of the silylmethyl species.

Step 2. Option D is too specific and **inconsistent with the drawing**: the structure shown is a **contact ion-paired, chelated K⁺ complex**, not clearly a solvent-separated species enforced by two DME ligands. → **Rejected D.**

Step 3. The boronate has electronic effects but is not the most direct element controlling regioselectivity change in DME. Steric blocking from Si groups alone does not explain it either. → Ruled out B and C.

Step 4. The **polyoxygenated/aza coordination environment around K⁺** changes the ion pair character, altering nucleophilicity and pathway of silyl transfer. → **Answer: A**

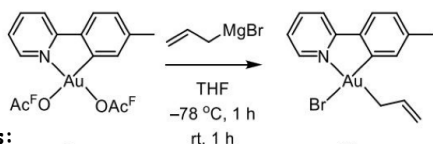
| **Perception X** | Knowledge — | Attribution — |

The model correctly identifies K⁺, the azine nitrogen, and surrounding oxygen donors, but misinterprets the ion-pairing state as a contact ion pair rather than a solvent-separated ion pair enforced by DME chelation. This error originates at the perception level: the model does not weigh option D and judge it less important, but instead explicitly rejects it as “inconsistent with the drawing,” indicating a misreading of the structural context rather than a failure to prioritize among correctly perceived features.

Figure 11: Perception error example. The model misidentifies or overlooks a structural feature visible in the image, leading to an incorrect answer.

Question:

In this reaction, the σ -allyl gold product forms with the allyl group arranged opposite the ring-based heteroatom binding site. Which structural feature in the structure shown most directly explains this preferred arrangement?

**Choices:**

- A. The cyclometalated C[^]N ligand framework enforces a rigid square-planar Au(III) geometry that defines a specific open coordination site during substitution, steering allyl incorporation into the position opposite the ring-based heteroatom binding site and yielding the observed trans arrangement.
- B. The bromide supplied under allylmagnesium bromide conditions binds strongly to Au(III) and reorganizes the ligand set during substitution, positioning the remaining incoming allyl group opposite the ring-based heteroatom binding site and forming the σ -allyl gold complex with the observed trans arrangement.
- C. The two fluorinated acetate (OAcF) ligands act as anionic, oxygen-bound ligands that undergo ligand exchange under Grignard conditions, and the resulting substitution pattern places the incoming allyl group opposite the ring-based heteroatom binding site, producing the σ -allyl complex with the observed trans arrangement.
- D. The pyridine-like N donor in the cyclometalated C[^]N framework exerts a trans influence that labilizes the Au–X bond trans to N, directing allyl substitution at that position and giving the σ -allyl complex with allyl trans to N.

GPT-5.4 Answer: A **Correct Answer: D**

GPT-5.4 Reasoning:

Step 1. The starting complex has a cyclometalated C[^]N ligand binding through a carbon and a pyridine-like nitrogen, giving a rigid square-planar Au(III) framework. Incoming substitution occurs at a defined site within this preorganized geometry.

Step 2. Option D's trans influence argument is tempting, but in cyclometalated Au(III) systems the stronger trans influence usually comes from the aryl carbon donor, not the pyridine-like nitrogen. So the N donor is not the decisive trans-influence element. → **Rejected D.**

Step 3. Bromide (B) and OAcF ligands (C) are participating/leaving groups, not the principal directors of geometry. → Ruled out B and C.

Step 4. The primary cause is the rigid square-planar organization imposed by the C[^]N ligand framework, which defines the substitution site. → **Answer: A**

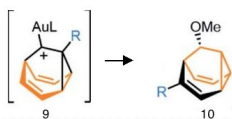
| Perception ✓ | Knowledge ✗ | Attribution — |

The model correctly identifies the square-planar Au(III) geometry and the C[^]N ligand framework. However, it applies an overgeneralized assumption that the aryl carbon donor dominates trans influence in cyclometalated Au(III) systems, and therefore dismisses the role of the pyridine-like nitrogen. In this system, substitution is directed at the position trans to the N donor, as supported by the literature. This misapplication of chemical knowledge leads the model to reject the correct mechanism (D) and instead select a generic geometric explanation (A).

Figure 12: Knowledge gap example. The model correctly perceives the structural features but lacks the chemical knowledge to connect them to the observed effect.

Question:

In this reaction, the methyl ether 10 is trapped and isolated as the major product instead of proceeding to the untrapped BB framework. Which structural feature in the structure shown most directly explains this outcome?

**Why Rejected**

The correct answer (C) identifies a feature visible in the image — the carbocation center in intermediate 9 — but the reasoning relies on a general organic reactivity principle: **nucleophilic attack on an electrophilic carbocation**. A reader recognizing this textbook-level concept can select C from the option text alone, without analyzing the depicted structure. In contrast, the three distractors invoke structure-specific reasoning (cage geometry, AuL electronic effects, and R-group stabilization) that requires visual inspection of the molecular framework. As a result, this item does not reliably test structure-based reasoning and is therefore excluded from the benchmark.

Choices:

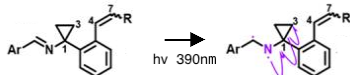
- A. The bridged barbaralane-type cage and embedded alkene unsaturation constrain intermediate 9 into a geometry that favors exo addition leading to C–O bond formation, locking in barbaralane methyl ether 10 rather than allowing formation of the untrapped BB framework.
- B. The AuL fragment bound to the cationic intermediate 9 increases carbocation-like electrophilicity at the reactive carbon, promoting rapid C–O bond formation that traps the cage as barbaralane methyl ether 10 instead of proceeding to the untrapped BB framework.
- C. The carbocation center (marked '+') in intermediate 9, positioned within the constrained cage framework, is highly electrophilic and readily undergoes nucleophilic attack by methanol, forming the C–O bond and trapping the system as methyl ether 10 instead of proceeding to the BB pathway.
- D. The R substituent on the cage framework stabilizes the developing positive charge in intermediate 9 and biases bond formation toward capture that installs OMe, producing barbaralane methyl ether 10 as the major product rather than the untrapped BB framework.

Correct Answer: C

Figure 13: Structure necessity rejection (76.7% of rejections). The correct answer can be selected without analyzing the depicted structure.

Question:

In this reaction, homolytic fragmentation occurs specifically at the bond labeled between the two numbered positions to generate a more favorable radical intermediate. Which structural feature in the structure shown most directly explains this selective bond cleavage?

**Why Rejected**

The labeled answer (B) attributes C1–C3 cleavage to **cyclopropane ring strain relief** — lowering the activation barrier for homolysis. However, option A offers a scientifically defensible alternative: the **fused aryl ring stabilizes the resulting radical through benzylic delocalization**, which lowers the energy of the product state and also promotes cleavage. In practice, both **thermodynamic driving force** (ring strain relief, B) and **product radical stabilization** (aryl delocalization, A) can serve as **primary explanations for the observed bond cleavage**. Because either explanation is chemically valid as the "most direct" cause, the question admits more than one defensible answer and is excluded from the benchmark.

Choices:

- A. The **aryl ring fused/attached to the cyclopropane (benzocyclopropane motif)** stabilizes developing radical character at the adjacent bond through benzylic radical stabilization via **conjugation with the aryl ring**; this stabilization promotes homolytic cleavage of the labeled C1–C3 bond, producing the observed fragmentation.
- B. The **cyclopropane ring (aminocyclopropane)** contains high inherent **ring strain** concentrated in the C1–C3 bond; relief of this strain lowers the barrier for homolytic C1–C3 scission, producing the observed fragmentation to a more favorable radical intermediate.
- C. The **alkynyl substituent (C≡C–R)** at the benzylic position withdraws electron density and stabilizes a nearby radical center through conjugative interaction with the aryl framework; this stabilization promotes homolytic cleavage of the labeled C1–C3 bond, producing the observed fragmentation.
- D. The **nitrogen adjacent to the cyclopropane (Ar–N connected to C1)** contributes to stabilization of the radical center via interaction with the nitrogen lone pair, lowering the energy of the product state; this stabilization can promote homolytic cleavage of the labeled C1–C3 bond to give the observed fragmentation.

Correct Answer: B

Figure 14: Answer unambiguity rejection (20.0% of rejections). More than one option is scientifically defensible.

```

► System
You are a chemistry expert decomposing literature explanations into causal components. Return only valid JSON.

► User
You are a chemistry expert analyzing a literature explanation about a chemical {molecule | reaction}.

RATIONALE FROM LITERATURE:
{rationale_sentence} [structure image attached]

YOUR TASK:
Decompose this rationale into its causal components AND identify other structural features visible in the attached image.

Step 1 – Extract the causal triple from the rationale
• EFFECT          What phenomenon or property is observed? e.g., "resists oxidation", "high regioselectivity", "requires Pd catalyst"
• STRUCTURAL_FEATURE The specific structural feature identified as the CAUSE. e.g., "tert-butyl groups at ortho/para positions", "C-Br bond in the substrate"
• MECHANISM         How does the feature cause the effect? e.g., "steric hindrance blocks oxidant approach", "oxidative addition requires transition metal"

Step 2 – Identify other visible features in the structure image
List 3-5 OTHER structural features that are clearly visible but are NOT the cause mentioned in the rationale. These will serve as plausible distractor sources when constructing the multiple-choice question in Stage 2.

For molecules: functional groups, ring systems, heteroatoms, substituent positions, stereochemistry, conjugation patterns
For reactions: other functional groups on substrates, leaving groups, protecting groups, ring systems, heteroatoms in the substrate or product

Step 3 – Classify the reasoning type
Molecules: "mechanism" | "design_rationale" | "structure_activity"
Reactions: "condition_rationale" | "catalyst_rationale" | "selectivity" | "mechanism" | "rate"

OUTPUT FORMAT (JSON only):
{
  "effect":          "concise description of the observed phenomenon or property",
  "structural_feature": "the specific structural feature that is the primary cause",
  "mechanism":       "how the feature causes the effect",
  "reasoning_type":  "one of the valid types listed above",
  "other_visible_features": ["feature_1", "feature_2", "feature_3"]
}

If the rationale is too vague to decompose (no clear structural feature or no clear causal mechanism), return: {"skip": true, "reason": "brief explanation"}

```

Figure 15: Prompt template for Stage 1 explanatory decomposition.

You are a chemistry education expert creating benchmark questions. Return only valid JSON.

You are constructing a chemistry benchmark question that tests **structure-grounded causal reasoning**. The question must be answerable only by someone who has examined this specific structure - general chemistry knowledge alone is not sufficient.

```

    Structural feature (CAUSE): {structural_feature}
    Mechanism: {mechanism}

```

OTHER FEATURES VISIBLE IN THE STRUCTURE (use for distractors):

{other visible features} [+ structure image attached]

1. Question must be short (1-2 sentences).
2. State ONLY the observed EFFECT – never name or hint at the responsible structural feature or its mechanism.
3. Ask "which structural feature most directly explains" or "what is the primary structural reason for" – framing that invites competition between multiple plausible features.
4. FORBIDDEN terms in question text (they reveal the structural cause or mechanism): structural descriptors (acceptor, donor, chelate, allylidenes); bonding/orbital terms (σ - π , charge transfer, resonance, hyperconjugation); mechanism terms (deprotonation, electrocyclic, tautomerization).
5. A reader armed only with general chemistry knowledge should not be able to reliably identify the correct option without seeing this specific structure.

Combine structural_feature + mechanism into one complete causal statement: "[feature] [mechanism] → [produces the stated effect]". Must be a faithful paraphrase of the literature – do NOT add information beyond what is stated in the feature and mechanism fields above.

CORE PRINCIPLE – Competing Causes in This Specific Structure: Every option must reference a feature from the OTHER FEATURES list AND propose a chemically valid mechanism for the same stated effect. The competition is about which feature is the primary cause – not whether it exists. Determining the dominant, most direct cause requires actually examining the molecular context, so general chemistry knowledge alone cannot resolve this competition.

Internal reasoning guide (stays in your head – never appears in option text): think about WHY each distractor is secondary (weaker effect, more indirect path, or supporting role in THIS molecule), but write the final option text without stating these reasons explicitly.

ALL options must explain the SAME stated effect. A distractor about a different effect (fluorescence, reactivity, stability) is trivially eliminable without seeing the structure.

Do NOT reference features absent from this structure. Every distractor must use a real feature from the OTHER FEATURES list.

FORBIDDEN hedging words in ANY option: "but", "though", "however", "indirect", "secondary", "supporting role", "weaker", "plays a role", "is proposed to", "would be expected to". Write ALL options with the same confident, direct, active-verb tone.

```
{
  "question": "1-2 sentence effect-only question",
  "correct_answer": "feature + mechanism + stated effect (no hedges, confident tone)",
  "distractors": [
    "other_feature_2 + mechanism + same stated effect (no hedges)",
    "other_feature_3 + mechanism + same stated effect (no hedges)",
    "other_feature_4 + mechanism + same stated effect (no hedges)"
  ],
  "reasoning_type": "mechanism | design_rationale | structure_activity | condition_rationale | catalyst_rationale | selectivity | rate"
}
```

29

► **System**
 You are an expert chemistry educator and assessment specialist. Your task is to evaluate the quality of a multiple-choice question designed to test structure-grounded causal reasoning in chemistry. Return only valid JSON.

► **User**

CONTEXT
 This question is part of the ChemCausal benchmark, which tests AI models' ability to perform **structure-grounded causal reasoning** – identifying why a specific molecular or reaction structure causes a particular chemical effect or property.

Key Design Principle – Same-Effect Competition: All four answer options explain the *same observed effect* but propose *different structural/mechanistic causes*. The distractors reference real structural features visible in the molecule/reaction – they are intentionally plausible. Only one option identifies the primary cause supported by the literature.

QUESTION TO EVALUATE:
 Question ID: {question_id} | Type: {molecule | reaction}
 Question: {question_text}
 Options: A: {...} B: {...} C: {...} D: {...}
 Rationale (literature source for correct answer): {rationale_sentence} [+ structure image]

EVALUATION DIMENSIONS – explain your reasoning before each score

- Distractor Quality (1-5) – Are the incorrect options challenging and plausible?**
 - 5 – All 3 wrong options are genuinely misleading; a trained chemist must carefully analyze this specific structure to rule them out.
 - 3 – Wrong options describe related chemistry but can be ruled out with general knowledge – no need to look at this specific structure.
 - 1 – Trivially wrong; factual errors or completely unrelated to the question.
- Question Meaningfulness (1-5) – Does this question probe a scientifically insightful causal relationship?**
 - 5 – Tests a scientifically insightful causal relationship specific to this structure; a chemistry expert would find the causal chain non-trivial and worth understanding.
 - 3 – Has some value, but the causal reasoning is overly simple or the question is too narrow.
 - 1 – No scientific value; tests only factual recall with no causal reasoning.
- Correct Answer Unambiguity (1-5) – Is the marked correct answer scientifically correct as the primary cause?**

Important: Do NOT penalize the correct answer simply because other options also sound reasonable – that is by design. Evaluate whether the marked answer accurately identifies the *primary* structural/mechanistic cause as supported by the literature rationale.

 - 5 – Clearly correct; structural analysis and the literature rationale strongly support this as the primary cause.
 - 3 – Plausibly correct, but there is genuine scientific debate about whether it is truly the primary cause of this effect.
 - 1 – The marked answer appears incorrect – another option is clearly a better explanation for the observed effect.
- Structure Necessity (1-5) – Must you analyze the structure image to answer accurately?**
 - 5 – Must see the structure; the answer depends on specific visual features (substituent positions, connectivity, stereochemistry) that are invisible from the question text alone.
 - 3 – Structure is helpful but not essential; general chemistry principles likely suffice to identify the correct option.
 - 1 – Fully answerable from the question text and options alone, without seeing the structure.
- Difficulty Prediction – Easy | Medium | Hard | PhD-level**
- Overall Assessment**
 - ACCEPT** – The question meets benchmark quality standards on all dimensions.
 - REVISE** – The question has merit but needs specific improvements before inclusion; explain what should be fixed.
 - REJECT** – Fundamental quality issues; should not be included without major revision; explain why.

OUTPUT FORMAT (JSON only):

```
{
  "question_id": "{question_id}", "evaluator_model": "{model_name}",
  "evaluation": {
    "distractor_quality": { "reasoning": "...", "score": 4 },
    "question_meaningfulness": { "reasoning": "...", "score": 5 },
    "correct_answer_unambiguity": { "reasoning": "...", "score": 5 },
    "structure_necessity": { "reasoning": "...", "score": 4 },
    "difficulty_prediction": { "reasoning": "...", "difficulty": "Hard" }
  },
  "overall_assessment": { "recommendation": "ACCEPT", "reasoning": "..." }
}
```

Figure 17: Prompt template for LLM-as-a-Judge quality screening on four dimensions.

ChemCausal Benchmark Evaluation - Evaluator

Introduction

Q4: PMC12663143_MOL_062_Q1

Question 4 of 600: PMC12663143_MOL_062_Q1

molecule

Chemical Structure

Question: This molecule is obtained in 61% yield with no loss of stereoselectivity and is used as a tyrosine-mimicking phenylalanine analog in medicinal chemistry. Which structural feature in the structure shown most directly explains this tyrosine-mimicking behavior?

Step 1 — Answer first: Select your answer below. The correct answer, rationale, and evaluation criteria will be revealed only after you submit.

A) The difluoromethyl (CF₂H) substituent on the aromatic ring polarizes its C–H bond through the two fluorines, enabling an OH-like hydrogen-bonding interaction while retaining lipophilicity, which produces tyrosine-mimicking behavior in binding sites.

B) The benzylic methylene linker between the aromatic ring and the backbone sets the aromatic ring at a specific distance and rotational freedom from the stereocenter, matching tyrosine-like spatial presentation in proteins and producing tyrosine-mimicking behavior.

C) The chiral alpha-carbon stereocenter in the amino acid backbone enforces a specific three-dimensional presentation of the aromatic side chain that matches tyrosine's preferred bound conformation in protein binding sites, producing tyrosine-mimicking behavior in medicinal chemistry.

D) The multiple carbonyl-containing functional groups (ester and carbamate C=O) along the backbone create an array of hydrogen-bond acceptor sites that collectively reproduce the polar interaction profile of tyrosine's phenolic oxygen, producing tyrosine-mimicking behavior in binding sites.

Submit Answer →

← Previous

Next →

Progress

Total Questions: 600

Completed: 0

Progress: 0%

Export CSV

Import CSV

Your Performance

Answered 3 / 600

Correct 0

0%

Data Storage:

Your evaluations are auto-saved to browser LocalStorage. They persist across sessions but will be lost if you clear browser cache.

Figure 18: Human evaluation platform: blind answer phase, in which experts answer the question without access to the source paper text.

31

ChemCausal Benchmark Evaluation - Evaluator

Introduction

Q4: PMC12663143_MOL_062_Q1

the polar interaction profile of tyrosine's phenolic oxygen, producing tyrosine-mimicking behavior in binding sites.

Literature Source Evidence

Direct difluoromethylation of phenylalanine-derived aryl bromide led to the formation of product 36 in 61% yield without erosion of the stereoselectivity; due to the capacity of the CF₂H group to serve as a lipophilic hydrogen bond donor, it could potentially function as a tyrosine analog in medicinal chemistry.

Click to View Evaluation Rubrics (Reference)

Your Evaluation

1. Distractor Quality

Are the incorrect options challenging and plausible?

Consider: Do distractors represent common misconceptions? Would a knowledgeable student need to think carefully to eliminate them?

1

2

3

4

5

Confidence:

Low

Medium

High

2. Question Meaningfulness

Does this question probe a scientifically insightful causal relationship?

Consider: Is the causal chain non-trivial? Would a chemistry expert find this question worth asking? Or is it testing obvious facts with no real causal depth?

1

2

3

4

5

Confidence:

Low

Medium

High

3. Correct Answer Unambiguity

Based on your chemistry expertise and the structure, is the marked answer scientifically correct as the primary cause?

Consider: Does the marked answer correctly identify the primary structural cause of the observed effect? Is the causal reasoning scientifically sound?

1

2

3

4

5

Confidence:

Low

Medium

High

Progress

Total Questions: 600

Completed: 0

Progress: 0%

Export CSV

Import CSV

Your Performance

Answered 4 / 600

Correct 0

0%

Data Storage:

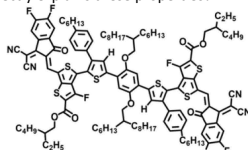
Your evaluations are auto-saved to browser LocalStorage. They persist across sessions but will be lost if you clear browser cache.

Figure 19: Human evaluation platform: evidence review and scoring phase, in which experts review the source paper evidence and score the four quality dimensions on a 1–5 Likert scale.

32

Question:

The structure shown exhibits a highly planar backbone and strong intramolecular charge transfer (ICT) character. Which structural feature most directly explains these properties?

**Choices:**

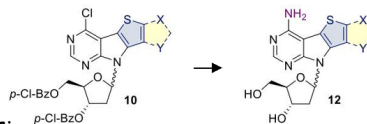
- A. Fluorinated aromatic rings introduce strong C-F bond polarization and aryl-fluorine steric/electrostatic effects that bias adjacent ring dihedral angles toward coplanarity and enhance electronic polarization, which yields a planar backbone with excellent intramolecular charge transfer.
- B. Multiple thiophene rings in the backbone extend the aromatic π -framework and promote quinoid character, enhancing delocalization and favoring a more coplanar arrangement of adjacent rings, which yields a planar backbone with excellent intramolecular charge transfer.
- C. Multiple intramolecular non-covalent conformational locks (S $\cdots$ O, H $\cdots$ O, and F $\cdots$ S contacts) along the conjugated backbone restrict interring torsion and enforce coplanarity, increasing orbital overlap and thereby producing a planar backbone with excellent intramolecular charge transfer.
- D. Terminal dicyano (-(CN)₂ / -CN) electron-withdrawing groups polarize the π -system through strong inductive and mesomeric effects, increasing backbone electronic polarization and stabilizing a delocalized planar geometry that yields a planar backbone with excellent intramolecular charge transfer.

Correct Answer: C

Figure 20: Mechanistic question on a molecule.

Question:

In this reaction, a halogen on the heteroaromatic core is replaced by an amino group and the sugar protecting groups are removed to give free hydroxyls in good yield. Which structural feature in the structure shown most directly explains this outcome?

**Choices:**

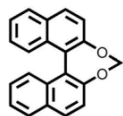
- A. The 4-chloro substituent on the fused electron-deficient heteroaromatic base is activated toward nucleophilic aromatic substitution by the electron-poor ring system; aqueous NH₃ displaces the chloride to install the amino group, and the same hot basic conditions cleave the p-chlorobenzoyl esters, replacing the halogen with an amino group and unmasking the sugar hydroxyls in good yield.
- B. The two electron-withdrawing ring nitrogens in the pyrimidine portion lower the LUMO across the fused heteroaromatic core and activate an unsubstituted ring carbon toward direct nucleophilic amination by aqueous NH₃, while the heated basic medium hydrolyzes the sugar esters, replacing the halogen with an amino group and unmasking the sugar hydroxyls in good yield.
- C. The sulfur atom in the fused thienyl ring raises electrophilicity of the entire heteroaromatic scaffold through d-orbital participation and increased polarizability, enabling aqueous NH₃ to displace the halogen, while the heated aqueous basic medium cleaves the sugar esters, replacing the halogen with an amino group and unmasking the sugar hydroxyls in good yield.
- D. The p-chlorobenzoyl ester protecting groups undergo preferential aminolysis by aqueous NH₃, generating a strongly basic alkoxide microenvironment at the sugar oxygens that activates the adjacent heteroaromatic system toward nucleophilic displacement, replacing the halogen with an amino group and unmasking the sugar hydroxyls in good yield.

Correct Answer: A

Figure 21: Mechanistic question on a reaction.

Question:

The structure shown displays a locked dihedral angle that increases molecular rigidity and improves chiral induction ability in coassemblies. Which structural feature most directly explains this effect?

**Choices:**

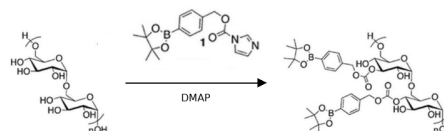
- A. The atropisomeric binaphthyl axis enforces a persistent axial chiral geometry by limiting rotation about the biaryl bond, locking the dihedral angle and producing increased rigidity with improved chiral induction ability.
- B. The two aryl-O- ether linkages rigidify the periphery through conformational restriction around the C-O bonds, locking the dihedral angle and producing increased rigidity with improved chiral induction ability.
- C. The methylene bridge (-CH₂-) linking the two oxygen atoms forms an O-CH₂-O tether that constrains rotation about the binaphthyl axis, locking the dihedral angle and producing increased rigidity with improved chiral induction ability.
- D. The binaphthyl polycyclic aromatic framework creates a sterically congested biaryl scaffold that raises the barrier to conformational interconversion, locking the dihedral angle and producing increased rigidity with improved chiral induction ability.

Correct Answer: C

Figure 22: Design question on a molecule.

Question:

Nanoparticles from the structure shown rapidly disassemble in phosphate buffer containing 1 mM H₂O₂ but remain stable without H₂O₂. Which structural feature most directly explains this H₂O₂-dependent disassembly behavior?

**Choices:**

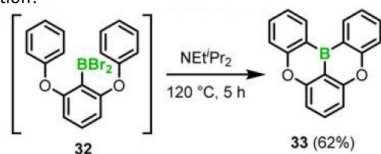
- A. The aromatic phenyl ring linker undergoes H₂O₂-mediated oxidative transformation to a more polar aromatic oxygenated product, reducing hydrophobic interactions between substituted segments and increasing water affinity, disrupting the assembled core and producing nanoparticle disassembly.
- B. The dextran polysaccharide backbone with many hydroxyl groups undergoes H₂O₂-driven oxidation that increases hydrophilicity and changes polymer-polymer association in buffer, producing a solubility shift that disrupts nanoparticle cohesion and causes disassembly.
- C. The arylboronic ester (B(pin)) groups appended to the dextran undergo selective oxidation by H₂O₂ to a more polar/ionizable boron/phenolic product, switching solubility and disrupting assembly-driving interactions, producing rapid nanoparticle disassembly.
- D. The p-nitrophenyl carbonate (activated carbonate) linker undergoes H₂O₂-promoted hydrolysis/cleavage in aqueous phosphate buffer to generate more hydrophilic fragments, breaking amphiphilic balance and disrupting intermolecular packing, producing rapid nanoparticle disassembly.

Correct Answer: C

Figure 23: Design question on a reaction.

Question:

This reaction proceeds without a Lewis acid activator (e.g., AlCl_3). Which structural feature in the structure shown most directly explains that observation?

**Choices:**

- A. The two aryl–O–aryl ether linkages donate electron density into the adjacent aromatic rings, increasing their nucleophilicity and enabling electrophilic activation/attack, which allows the transformation to proceed without a Lewis acid such as AlCl_3 .
- B. The triarylboron center formed in the product stabilizes through strong B–C bond formation and delocalized aryl–boron interaction, making product formation highly thermodynamically favorable and allowing the transformation to proceed without a Lewis acid such as AlCl_3 .
- C. The BBr_2 -substituted organoboron unit increases the intrinsic electrophilicity at boron and promotes intramolecular aryl transfer to boron under heating, which drives triarylboron formation and allows the reaction to proceed without a Lewis acid such as AlCl_3 .
- D. The polycyclic aryl framework provides a rigid, preorganized geometry that enforces close proximity of the reacting aryl sites to the boron center, accelerating intramolecular bond formation and allowing the transformation to proceed without a Lewis acid such as AlCl_3 .

Correct Answer: A

Figure 24: Condition question on a reaction.